

Novel Methods for Visualising Graphs



Osman Akbulut

School of Computing
Newcastle University

Submitted in Fulfilment of the Requirements for the Degree of
Doctor of Philosophy

April 2024

To my family

Declaration

“I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 80,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.”

Osman Akbulut
April 2024

Acknowledgements

I would like to extend my utmost gratitude to all who have afforded me the opportunity to successfully finalise this dissertation. First, I would like to thank my wonderful supervisors, **Professor Nicolas Holliman** and **Dr Matthew Forshaw**, for their continued guidance and support throughout such uncertain and challenging times at the university. Without this, it would not have been possible to complete the study. Their helpfulness and encouragement made the process very rewarding. I would also like to thank **Dr Tong Xin**, a member of my supervisory team, who supported me with her valuable comments and suggestions in the course of my study.

I would like to acknowledge the contributions of Lucy McLaughlin and appreciate her dedication and input in this study during our period of collaboration.

In addition, I would like to express my sincere thanks to **the Ministry of National Education of the Turkish Republic** for giving me the scholarship to meet the expenses and pursue my doctorate and master of science studies in the United Kingdom.

Special thanks go to my friends Dr Kenan Koc, Dr Mutlu Yapici, and Burak Kucukgoz for sharing their time and experience.

I would like to express my profound appreciation to my father and beloved sister for their unwavering affection, consistent assistance, and motivational influence that have been instrumental in shaping my life. It would not have been possible without you all by my side.

Furthermore, I want to dedicate this dissertation to my wife, Esra, for supporting me in everything I did and loving me unconditionally. Without her support, I would have never been able to accomplish this work.

Finally, I would like to open an extra paragraph for my beloved mother. I know that words are not enough to describe her. I cannot express how much she means to me. I always feel her presence, even when fate separates us too soon. Thank you so much, Mom, for everything you did for me. I want to dedicate this thesis to you. I will always love you.

Abstract

Graphs/networks are fundamental mathematical constructs that play a crucial role in representing and analysing diverse real-world phenomena across various disciplines. However, rapidly increasing quantities of data pose significant challenges for informed decision-making. Visual clutter increases proportionately with the graph's size and complexity, obscuring semantic relationships and limiting human comprehension. Additionally, graph visualisation research has primarily focused on depicting graphs based on their primary values without considering the uncertainty inherent in the data. This could yield visual representations that lead to overlooking unrevealed trends and patterns or misinterpretations of the underlying data by human viewers. Consequently, there is an increasing need for methodologies that assist end users in understanding their data and its inherent structure, thereby facilitating an effective analysis and better decision-making procedure under uncertainty.

This thesis outlines two research objectives:

- Addressing graph summarisation challenges through proposing a summarisation algorithm.

The first objective is to understand the network data pertaining to links and nodes, with less emphasis on the network's structure and connectivity. Users require simplified visualisations that clearly convey the relationship between network structure and associated data. We developed an algorithm to summarise graph-based data by extracting the maximum information in readable and informative forms of the original graph to end-users.

- Exploring the development of a novel visualisation approach to aid the design, implementation, and operation of visual search tasks on node-link diagrams.

The second objective is to identify and address existing approaches' limitations and challenges. This research introduces a node-link visualisation model designed for visually representing and analysing bivariate networks. We demonstrate it effectively addresses the challenges associated with these approaches. The major contributions of the thesis are as follows:

1. We present a novel node-link visual model — visual entropy (Vizent) graph — to effectively represent both primary and secondary values, such as uncertainty, on the edges simultaneously.
2. We present the novel Vizent edge design and empirically demonstrate (in collaboration with Lucy McLaughlin) that different edge glyphs have a perceived order through pairwise testing.

3. We perform two task-based usability studies to demonstrate the efficiency and effectiveness of our approach for visualising bivariate networks using static node-link diagrams.
4. We compare the Vizent design against three visual encodings selected from the literature on various graphs ranging in complexity from 5 to 25 edges for three different tasks.

Keywords: Information Visualisation, Graph Visualisation, Graph Summarisation, Node-Link Diagram, Edge Visualisation

Table of contents

Acknowledgements	vii
Abstract	ix
List of figures	xv
List of Algorithms	xxi
List of tables	xxiii
1 Introduction	1
1.1 Information Visualisation	2
1.1.1 Understanding Uncertainty in Data	5
1.1.2 Challenges of Visualising Uncertainty	5
1.2 Graph Drawing and Graph Visualisations	6
1.3 Motivation and Research Questions	8
1.4 Contributions	10
1.5 Outline	11
1.6 Related Publication	11
2 Background	13
2.1 Human Perception	14
2.2 Graph	17
2.2.1 Main Graph Definitions	17
2.2.2 What is a Graph?	19
2.2.3 Graph Visualisation	22
2.2.4 Graph Layout	28
2.3 Visualisation of multivariate network	32
2.4 Uncertainty Visualisation	39
2.4.1 Uncertainty	39
2.4.2 Uncertainty Visualisation	40
2.4.3 Uncertainty Visualisation in Graphs	43
2.4.4 Summary	45
2.5 Graph Summarisation	46
2.5.1 Summary	50

3	Graph Summarisation Method	53
3.1	Introduction	54
3.2	Motivation	55
3.3	Facebook Datacentre Topology	56
3.3.1	Facebook’s Fabric Network Design	57
3.3.2	Traffic pattern inside Facebook’s datacentre network	57
3.4	Concept of the Variance in edge weights	59
3.4.1	Edge Variance	59
3.4.2	Applying Edge Variance Concept to Facebook Traffic Patterns	60
3.5	Information Theory	61
3.5.1	Shannon Entropy	61
3.6	Entropy-Based Graph Summarisation Algorithm	64
3.6.1	Overview	64
3.6.2	Implementation	65
3.7	Testing with Facebook network data	69
3.8	Limitations	73
3.9	Future work	73
3.10	Summary	74
4	Introduction Vizent Edges	77
4.1	Introduction	78
4.2	Motivation	78
4.3	Related Works	79
4.4	Glyph-based Visual Design	80
4.5	Introduction To Vizent Edge Design	81
4.5.1	Visual Entropy (Vizent) Edge Design	83
4.5.2	Null Case Representation	83
4.5.3	Experimental Process of the Edge Glyph	84
4.6	Results	87
4.6.1	Effect Size	88
4.7	Summary	90
5	A novel node-link visual model – visual entropy (Vizent) graph	91
5.1	Introduction	92
5.2	Experiment 1: Edge Performance Experiment	93
5.2.1	Usability Study of Bivariate Network Visualisation Approaches	93
5.2.2	Control Group of Visual Encodings	94
5.2.3	Tasks	98
5.2.4	Hypotheses	99
5.2.5	Experimental Procedure	99
5.2.6	Edge Performance Experiment Results	100
5.2.7	Graph Size	105

5.2.8	Task Type	109
5.3	Experiment 2: Node and Edge Performance Experiment	110
5.3.1	Hypotheses	113
5.3.2	Experimental Results	113
5.3.3	Participants' comments on the experiment	115
5.4	Discussion and Conclusion	115
5.5	Limitations	117
5.6	Future Works	117
6	Conclusion	119
6.1	Summary of Contributions	122
6.2	Answers to research questions	123
6.3	Future Work	124
6.3.1	Generalisation	125
	Bibliography	127
	Appendix A Experiment 1	141
	Appendix B Experiment-2	149

List of figures

1.1	The information visualisation reference pipeline [33]. (Source: [209], p. 99) . . .	3
1.2	The four basic dataset types by Munzner [133]: Tables, Networks, Fields, Geometry (Spatial).	4
1.3	The attribute types by Munzner [133].	4
1.4	Two visual representations of the same network.	7
2.1	In the study by Ware [201] provided a comprehensive list of pre-attentive examples with a broad explanation, yet some of the pre-attentive features were given.	15
2.2	Two examples of pre-attentive processing: spotting a red object from a set of black ones and spotting one circle from a set of red squares.	15
2.3	The figure represents a subset of Gestalt Laws (or Principles) [110].	16
2.4	An example of different types of graphs.	18
2.5	A visual representation of the graph G	19
2.6	An example of a multigraph with multiple orange-coloured edges. [26]	20
2.7	Illustrating a bipartite graph that has two node types. [26]	21
2.8	Illustrating the representation of a rooted tree utilising the <i>TreeLayout</i> algorithm [37].	21
2.9	Node-link graph drawings of Königsberg problem	22
2.10	Bertin [23] proposed various visual encodings for representing graphs.	23
2.11	The diagram illustrates the procedure for converting a pseudograph into a graph employing a bipartite layout, which involves duplicating the vertices and arranging them equidistantly along one axis of two parallel vertical axes. The edges are then represented as straight links connecting the vertices on the two axes. [20]	24
2.12	Traditional tree-based visualisation techniques for hierarchical data in which colour encodes to univariate categorical data. Treemaps (a), radial trees (b), icicle plots (c), sunburst charts (d), and circular treemaps (e). [213]	25
2.13	Illustration of a Treemap layout that depicts housing data pertaining to several boroughs of London. Attributes are represented by the encoding of size, colour, labels, and an approximation of position. [175].	25
2.14	Example of the different visual representations of static graphs. All representations show the same dataset [19]. Left three images: node-link diagrams. Right: Adjacency matrices (matrix).	26

2.15	Evaluation of node-link diagrams and adjacency matrix visualisations [67].	27
2.16	Hybrid graph visualisation combining node-link diagrams with adjacency matrices [77].	27
2.17	Random layout [160]	29
2.18	a) Circular layout employing edge-length minimising order. b) Circular layout employing exterior routing with edge clustering algorithm [65].	29
2.19	a) Illustration of a force-directed layout algorithm applied to a graph consisting of 75 nodes. The layout incorporates size encoding to represent edge properties. b) A larger network where the node size encodes node attributes.	30
2.20	Illustration of a graph visualisation on a map with geographic location [28].	31
2.21	The evaluation of colour encoding on nodes was conducted in node-link diagrams [98].	33
2.22	Three different visual mappings were evaluated on nodes in the context of node-link diagrams [66].	34
2.23	A example of node-link diagram maps attributes on nodes and edges using colour and size visual channels [2].	34
2.24	A node-link representation where nodes were placed according to their natural geographic location. Thickness and colour of edges encode to visualise network traffic [98].	35
2.25	a) Node size and colour were used to visually represent a categorical and a quantitative attribute, respectively. b) Multivariate data on nodes is represented by nested bar charts for numerical values, and coloured glyphs were used for categorical values [136].	36
2.26	Several numerical attributes are encoded by the width of the coloured segments [107].	36
2.27	Link using coloured bars for encoding four quantitative edge attributes. This approach involves assigning specific colours to each attribute and representing their values through the length or height of the corresponding coloured bar. [161].	37
2.28	The visualisation pipeline can be extended to incorporate considering uncertainty at each stage. (Source: [205], p. 3))	41
2.29	Eleven visual channels have been examined for representing uncertainty on point symbols [122].	42
2.30	To determine which visual designs individuals preferred concerning uncertainty encoding, they requested that participants choose one of four possible options [27].	43
2.31	Guo et al. [71] conducted a user study of paired visual variables to visualise uncertainty as an additional dimension in conjunction with the primary attribute of the edge.	44
2.32	Eight different combinations of visual variables used in their experiment [13].	45
2.33	The layout of the Internet graph (top) has a total of 149661 edges. After implementing the edge bundling method, the resulting image is at the bottom; however, the result is still a hairball cluttered with edges. [64].	47

2.34	The example of graph-structure summarisation [114] is centred on graph topologies and does not incorporate attribute values.	48
2.35	a) The original network is visualised using the node degree information. b) The graph is simplified through the removal of structural information, namely by eliminating one-degree nodes in order to reduce the visual complexity. [169]. . .	49
3.1	Fabric network design of Facebook data centre, which accommodates a large number of interconnected devices and a more complex topology [5].	56
3.2	Illustration of a Server Pod which has only 48 server racks (Top-of-rack Switches) [5].	57
3.3	“force-directed” layout algorithm is used within the visualisation tool for representing 245349 distinct source IPs (server machine) and 116152 distinct destination IPs. The graph was zoomed in on a particular section of a larger graph. (Source: Author’s own)	58
3.4	The network was shown as a graph with server racks being nodes and the variance of bandwidth being flows on edges. (Source: Author’s own)	60
3.5	Plot of binary entropy used by Shannon in [166].	63
3.6	A screenshot from the study of Karloff and Shirley showing the 56-node summary tree of the math genealogy, utilising the maximum entropy approach. The colours of nodes are determined by their depth-1 ancestor, while the sizes of nodes are proportional to their weights in the summary tree. [99]	65
3.7	The flowchart outlines the formation of separate edges, commencing from a minimum of two edges and continues until the desired number, as specified by the user, is reached.	67
3.8	A screenshot of the 4-edged graph summary, which the proposed algorithm generates results in JSON (Source: Author’s own figure)	69
3.9	Figures present the relationship between execution time and memory usage of the algorithm when processing Facebook network data with differing complexities—1500 edges in (a) and 1000 edges in (b).	70
3.10	Before implementing the proposed approach, the Facebook network data was visualised as a graph with servers being nodes and the variance of packet length being flows on edges (Source: Author’s own)	71
3.11	For a given number of $n = 10$ edges, the resulting summary graph $S = (V, E)$ consists of the collection of sets of distinct edges comprising $(n - 1)$ selected distinct edges and one <i>superedge</i> . Each generated summary subgraph exhibits maximum information given the same number of edges from an information theoretical point of view.	72
4.1	A Various Range of Visual Channels [35]. (Source: [25])	81
4.2	Holliman et al. [80] proposed a novel set of visual entropy glyphs. The value increases from left to right proportionally to the complexity (frequency) of the shape.	82
4.3	An example of a new node-link visual model. Visual Entropy (Vizent) edge design combining colour and stripe pattern.	83

4.4	This edge glyph design is developed for cases where the dataset exhibits instances of missing uncertainty value.	84
4.5	The novel design of the Visual Entropy (Vizent) edge glyphs. In the experiment, every pair of a set of seven edge glyphs were shown for comparison.	85
4.6	The figure shows the instructions that were presented to each participant prior to the commencement of the trial.	86
4.7	An example of a pairwise comparison image used in the trial.	86
4.8	The results show the probability of correct response with 95% confidence intervals for each glyph, (n=27).	88
4.9	Mean response time for each glyph, (n=27).	89
5.1	Illustrating practice examples of node-link visualisations used in the practice trial. Vizent design (a) and three different visual encodings, (b), (c), and (d), were employed for comparison in our experiment. The graphs provided above have two attributes, each of which is mapped to different visual attributes. . . .	94
5.2	Default Edge Visual Variables of the Experiment.	95
5.3	A 10-edges graph used as stimuli of Vizent	95
5.4	A 10-edges graph used as stimuli of <i>Num</i>	96
5.5	A 10-edges graph used as stimuli of <i>Wid–Lig</i>	96
5.6	A 10-edges graph used as stimuli of <i>Sat–Tra</i>	97
5.7	The node-link diagram shows an example of a 25-edges graph used as stimuli of <i>Sat–Tra</i>	97
5.8	Response times in seconds (upper side, as box plots) and accuracy (bottom side, as bars representing mean values and error bars representing 95% confidence intervals) classified by Vizent, <i>Num</i> , <i>Wid–Lig</i> , <i>Sat–Tra</i> , respectively grouped by <i>Task</i> . The black lines highlight significance between visual encodings. . . .	102
5.9	Response times in seconds (upper side, as box plots) and accuracy (bottom side, as bars representing means and error bars representing 95% confidence intervals) classified by edge number, respectively grouped by <i>Task</i> within Vizent.	105
5.10	Response times in seconds (upper side, as box plots) and accuracy (bottom side, as bars representing means and error bars representing 95% confidence intervals) classified by edge number, respectively grouped by <i>Task</i> within <i>Num</i>	106
5.11	Response times in seconds (upper side, as box plots) and accuracy (bottom side, as bars representing means and error bars representing 95% confidence intervals) classified by edge number, respectively grouped by <i>Task</i> within <i>Wid–Lig</i>	107
5.12	Response times in seconds (upper side, as box plots) and accuracy (bottom side, as bars representing means and error bars representing 95% confidence intervals) classified by edge number, respectively grouped by <i>Task</i> within <i>Sat–Tra</i>	108
5.13	Visual entropy glyphs encode hourly mean temperature values using the Met Office colour scale and the variance of those values in the urban digital twin application context [80].	110

5.14	Example of the Vizent graph design. Node variables are mapped to Vizent glyph shapes.	111
5.15	Illustration of the Vizent graph design. Four encoding levels for network traffic value and its variability were assigned to Vizent edges and Vizent glyphs [80]. Each edge and node represent the combinations of a network traffic level (20 to 80) and its variability level (2 to 8).	112
5.16	Illustration of <i>Num</i> visualisation. Four encoding levels for network traffic value and its variability were represented by numerical values. Each edge and node represent the combinations of a network traffic level (20 to 80) and its variability level (2 to 8).	112
5.17	Showing the follow-up experiment results. Mean response times in seconds (left side, as box plots) and accuracy (right side, as bars representing means and error bars 95% confidence intervals) classified by Vizent and <i>Num</i> graph representations, (n=24).	114
A.1	A 15-edges graph used as stimuli of <i>Num</i> for Task 1	141
A.2	A 15-edges graph used as stimuli of Vizent for Task 1	142
A.3	A 15-edges graph used as stimuli of <i>Wid–Lig</i> for Task 1	142
A.4	A 15-edges graph used as stimuli of <i>Sat–Tra</i> for Task 1	143
A.5	A 20-edges graph used as stimuli of <i>Num</i> for Task 2	143
A.6	A 20-edges graph used as stimuli of Vizent for Task 2	144
A.7	A 20-edges graph used as stimuli of <i>Wid–Lig</i> for Task 2	144
A.8	A 20-edges graph used as stimuli of <i>Sat–Tra</i> for Task 2	145
A.9	A 25-edges graph used as stimuli of <i>Num</i> for Task 3	145
A.10	A 25-edges graph used as stimuli of Vizent for Task 3	146
A.11	A 25-edges graph used as stimuli of <i>Wid–Lig</i> for Task 3	146
A.12	A 25-edges graph used as stimuli of <i>Sat–Tra</i> for Task 3	147
B.1	An example of the Numerical values visualisation used as a stimulus in the experiment.	149
B.2	An example of the Numerical values visualisation used as a stimulus in the experiment.	150
B.3	An example of the Numerical values visualisation used as a stimulus in the experiment.	150
B.4	An example of the Numerical values visualisation used as a stimulus in the experiment.	151
B.5	An example of the Numerical values visualisation used as a stimulus in the experiment.	151
B.6	An example of the Vizent graph used as a stimulus in the experiment.	152
B.7	An example of the Vizent graph used as a stimulus in the experiment.	152
B.8	An example of the Vizent graph used as a stimulus in the experiment.	153
B.9	An example of the Vizent graph used as a stimulus in the experiment.	153
B.10	An example of the Vizent graph used as a stimulus in the experiment.	154

List of Algorithms

1	Graph Summarisation Algorithm for a directed or undirected Graphs	68
---	---	----

List of tables

4.1	The table illustrates the edge glyph pairwise order comparisons alongside the outcomes of the exact binomial test for each individual glyph.	87
4.2	Interpretation of Cohen’s g effect size.	89
4.3	Effect Size Estimate Using Cohen’s g For Each Of The Edge Glyph Binomial Tests	90
5.1	List of experimental questions used in this study.	98
5.2	Post hoc analysis results with Wilcoxon signed rank tests for each <i>Task</i> . Significant differences between pairs of visual encodings are highlighted in bold using $p < 0.0125$	101
5.3	The mean and median response times in seconds are grouped by visual encoding and <i>Task</i> . The lowest mean response time of the visual encodings is highlighted in bold for each <i>Task</i> . The terms <i>Single</i> and <i>Dual</i> target visual search are also displayed.	103
5.4	The mean and median accuracy (%) are grouped by visual encoding and <i>Task</i> . The highest mean accuracy percentage of the visual encodings is highlighted in bold for each <i>Task</i> . The terms <i>Single</i> and <i>Dual</i> target visual search are also displayed.	104

Chapter 1

Introduction

Contents

1.1	Information Visualisation	2
1.1.1	Understanding Uncertainty in Data	5
1.1.2	Challenges of Visualising Uncertainty	5
1.2	Graph Drawing and Graph Visualisations	6
1.3	Motivation and Research Questions	8
1.4	Contributions	10
1.5	Outline	11
1.6	Related Publication	11

1.1 Information Visualisation

*“Data visualisation is not your creative outlet;
data visualisation is making data understandable.”*

— OpenVis Conf

Throughout history, visual representations have played a significant role in facilitating human reasoning and enhancing communication. Since ancient times, humans have employed various visual mediums, such as drawings and carvings, to communicate information and ideas. The visualisation field underwent a significant transformation in the late 20th century due to the advent of computer technology. In recognition of this burgeoning discipline, the first IEEE Visualisation Conference (IEEE VIS) was organised in 1990. Subsequently, visualisation techniques have persistently broadened in both academic and industrial domains, resulting in significant advancements and influence. In the last thirty years, visualisation has established itself as a progressively influential interdisciplinary domain, benefiting from contributions from various fields such as computer science, cognitive psychology, graphic design, statistics, and other related disciplines.

Information visualisation has emerged as an influential approach for transforming large, complex data into intelligible graphical representations, facilitating user insight and understanding of the underlying information [177]. A key functional role of information visualisation is enabling efficient and effective "visual information seeking" [172], allowing users to navigate and analyse the information space. As pioneering computer scientist Ben Shneiderman noted, "The purpose of visualisation is insight, not pictures." Put differently, the fundamental objective of visualisation is to facilitate comprehension and exploration rather than solely producing visually pleasing images.

Visualisation refers to the graphical depiction of data or concepts [101] while supporting various cognitive processes [31, 102, 162]. Nowadays, even end-users expect to be able to explore and understand the results. Visualisations can serve multiple purposes. These purposes include conducting exploratory analysis to identify patterns or anomalies, validating hypotheses through either a data exploration process or quantitative analysis, and effectively communicating the analysis findings to an audience.

- *Exploratory analysis.* The end user lacks specific and well-defined objectives or questions when engaging with the visual representation. In contrast, open-ended data exploration aims to identify new patterns, trends, outliers, and other discoveries within the information space. [57, 185]. The objective is to gain an understanding of the data, extract pertinent information, and formulate hypotheses.
- *Confirmative analysis.* The end user holds predefined hypotheses or questions to be confirmed by examining the data. The purpose of visualisation is to validate or invalidate hypotheses that may have been derived from data exploration or models that are linked to the data. [57, 102].

- *Descriptive or Presentation.* The end users already possess an understanding of the key phenomena and relationships within the data. However, they require effective visual representations and metaphors to convey these results and insights yielded by an exploratory analysis or analytical analysis process to others. This may include decision-makers, domain experts, or the general public, among other potential audiences. [185].

During the process of visualisation development, researchers analyse the dataset at hand while considering the desired objectives of end-users, intending to communicate information effectively through the visual display. Current research in data visualisation employs analytical methods based on computer graphics to produce visual depictions of information. Through transforming raw data into visual representations, visualisation techniques empower the rapid identification of patterns, relationships, and anomalies that might be challenging to perceive in the original, unprocessed data. This approach facilitates the acquisition of valuable insights and supports decision-making processes based on data analysis.

Figure 1.1 illustrates the visualisation reference pipeline commonly adopted in most information visualization systems as a base. The data model consists of four successive phases: *data source*, *data tables*, *visual abstraction* and *views*.

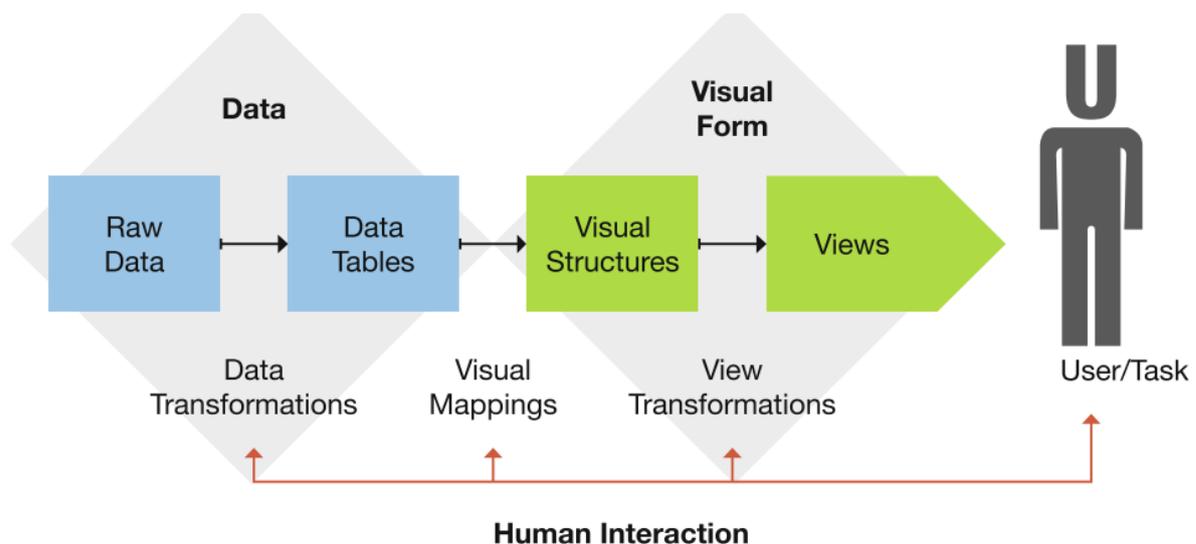


Figure 1.1 The information visualisation reference pipeline [33]. (Source: [209], p. 99)

Data transformation involves the initial stage of converting raw data into a format suitable for analysis through methods such as data aggregation or data filtering.

Visual mapping involves mapping transformed data into visual structures through encoding channels, such as colour and shape. This phase is of significant interest to designers, as it predominantly establishes the level of expressiveness and effectiveness of the resulting visualisation.

View transformation involves the generation of a visual output that incorporates the encoded data and presents it in the form of images or animations on a display monitor. The meticulous choice of rendering techniques is of utmost importance in the creation of significant final visual representations. In general, this sequential procedure offers a structure for converting unprocessed data into informative visual representations customised to meet the requirements of the intended audience.

As illustrated in Figure 1.1, a widely used framework facilitates the transformation of tabular datasets into visualisations through a series of sequential processing stages. The pipeline approach discussed is applicable and frequently used with simple formatting data (see Figure 1.2 for the different types of datasets). However, in practical situations, datasets tend to be complex and diverse, exhibiting non-linear structures. For example, transportation networks or organisational hierarchies demonstrate specific and important characteristics, which have a connection and containment between them by their nature.

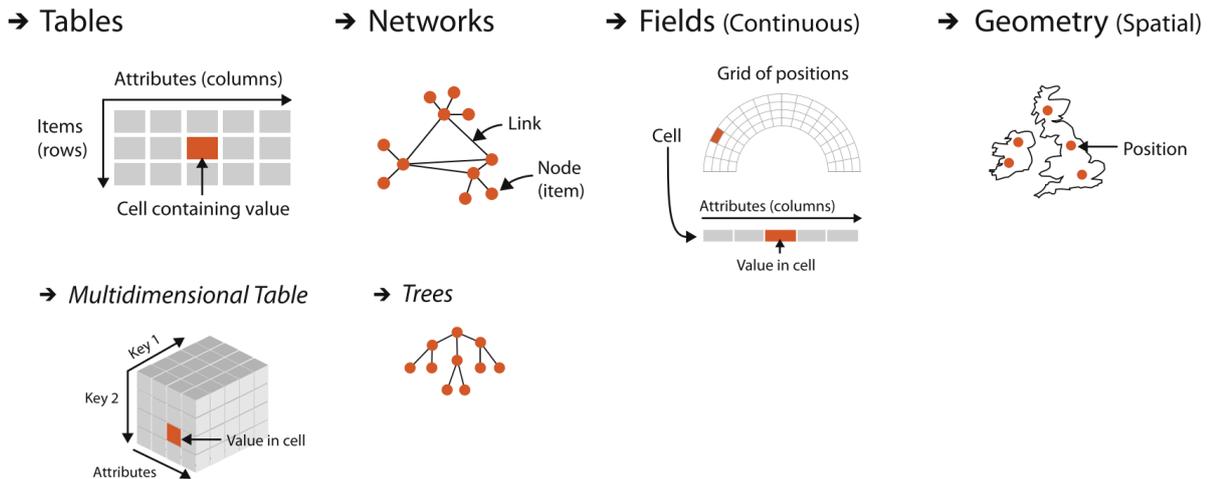


Figure 1.2 The four basic dataset types by Munzner [133]: Tables, Networks, Fields, Geometry (Spatial).

→ Categorical



→ Ordered

→ Ordinal



→ Quantitative



Figure 1.3 The attribute types by Munzner [133].

In the context of Information Visualisation (InfoVis), the objects within the network are commonly referred to as *nodes* or *vertices*, and the relationships between the objects are denoted as *links* or *edges*. The network itself is typically referred to as a *graph*. Additionally, the data associated with the objects and connections are commonly referred to as *attributes*, *features*, *dimensions*, or *properties* [104]. Data attributes are commonly categorised into different types, and multiple categorisation taxonomies can be found in visualisation literature. An attribute denotes a particular characteristic that is amenable to quantification, observation, or logging [133]. As shown in Figure 1.3, the main distinction of attributes is between *categorical* and *ordered* [34, 133], and the ordered classification is further subdivided into ordinal and quantitative categories. Additionally, the quantitative category is subdivided into *interval* and *ratio* data.

It should be noted that graphs can be designated, whereby both nodes and edges can have a set of associated attributes, which can be either quantitative or qualitative variables. For example, in an urban transit network, the fundamental objects consist of bus stops and train stations, which can be regarded as nodes, interconnected by routes, which can be considered as links. The dataset also encompasses information on the travel time between two stops or train stations and estimates of latency periods, distance, and number of people, all of which can be represented as attributes within the graph. The characteristics of such data types are paramount in understanding how the dataset is structured and often constitute the most vital elements that need to be communicated through visual depictions. The predominant visual representation of these relational data sets is node-link diagrams (graphs or networks), in which nodes symbolise specific elements (or objects), and links symbolise various interconnections or relations between these elements. Graphs convey the inherent characteristics of relationships (or connections) and enclosure properties, whereas tabular representations do not offer the same clarity or accessibility in conveying such information.

1.1.1 Understanding Uncertainty in Data

Uncertainty is a common issue when dealing with data; it simply means we are not always completely certain about the information at hand. Consider a weather forecast as an example. Even though it predicts when and where it might rain, there is still a degree of uncertainty. Similarly, when we look at graphs illustrating average travel times between train stations, it is important to recognise that these times can vary. Therefore, uncertainty value should be taken into account when analysing and communicating visual data. This can be achieved by simultaneously displaying the data value we possess and the degree of uncertainty.

1.1.2 Challenges of Visualising Uncertainty

Graph visualisation faces a significant challenge due to the inherent uncertainty in data, which arises from measurement errors, accuracy, missing values, or ambiguous definitions. This uncertainty can seriously mislead the interpretation of these visualisations. For example, illustrating only average travel times without considering the uncertainty through traditional methods may lead people to perceive these times as constant. However, factors like traffic, weather, or unforeseen delays can affect travel times, adding variability to the dataset. By incorporating uncertainty

data, the visual representations strive to provide a more accurate and comprehensive view of the underlying data structures.

Because of their limited capacity to accurately represent uncertainty data, current methods for visualising uncertainty often fail to facilitate decision-making processes. Therefore, this thesis aims to provide novel methods for effectively handling and visualising data value and its uncertainty within graph visualisations. This enhancement makes the visualisations more interpretable and supports better decision-making by recognising the variability of the data.

The following sections will present a general understanding of graph drawing and graph visualisations without delving into more details. (Section 1.2) and focus on the crucial role of representing uncertainties to facilitate well-informed decision-making (Section 1.3).

1.2 Graph Drawing and Graph Visualisations

In the era of big data, graphs (also known as networks) have become pervasive in diverse domains, including social networks, intelligence analysis, and network traffic. The terms graph and network are used interchangeably throughout the thesis. Graphs are a versatile tool for representing and analysing various complex systems that exist in our environment. These systems encompass a wide range of domains, including but not limited to internet networks, telecommunication networks, road networks, chemical compounds, social networks, genetic information, and power grids. Graphs facilitate the exploration of relationships among entities, hence enabling the acquisition of useful insights into the systems under investigation.

Graph Visualisation and *Graph Drawing* are reliant on the visual representation of graphs and have become frequently studied disciplines of research [197, 19]. *Graph Drawing* pertains to the study of graph *readability*, with specific emphasis on the presentation of node-link diagrams. In such contexts, the readability challenges are articulated through *aesthetic criteria*, which can be defined as aesthetic optimisation objectives for (static) drawing graphs [148]. To ensure the readability of graph drawings, the most common and significant aesthetic criteria include minimising edge crossings, uniform edge lengths, even node distribution, edge bend minimisation, avoiding node overlap, symmetry, and drawing area [203]. While these criteria are widely acknowledged, they do not hold equivalent significance. For example, Purchase et al. [148] demonstrate that “reducing the crossings is by far the most important aesthetic while the impact of minimising the number of bends and maximising symmetry is comparatively less significant”. Additionally, they do possess certain limitations. For example, these criteria cannot be met in the context of large graphs.

Graph structure evolving over time, where nodes and edges are added or deleted, is called a *dynamic* graph or *time-evolving* graph in contrast to static graphs. Time-dependent changes may affect the attributes of nodes and edges (such as the weights of the edges change), the graph structure itself (through adding or removing nodes and edges), or both. This type of network is typically represented as a sequence of single depictions, each corresponding to a certain timestamp. Many approaches use animated diagrams to show the changes, for example [60].

Beck et al. [18] conducted a study that built upon prior research by examining both static and dynamic graphs, regardless of their graphical representations. This includes matrix repre-

sentations as well as node-link diagrams. The researchers examined three categories of criteria: general scalability, dynamic scalability, and aesthetic scalability. The *general criteria* encompass reducing visual clutter, mitigation of spatial misunderstanding arising from close proximity, optimisation of spatial alignment for navigating paths, and maximising the efficient use of space. In the context of dynamic graphs, the *dynamic criteria* include maximising the stability of the displayed information between different time points, reducing the cognitive burden required for examining temporal dynamics, and minimising temporal aliases that may arise due to the positioning of multiple nodes in the same location across two distinct time periods. The *aesthetic scalability* criteria pertain to the readability of graphs when dealing with larger graphs. This includes scalability regarding the number of vertices, edges, and graphs, particularly as the number of time steps for which graph data is provided increases. All these criteria are important, but it is not possible to optimise them at the same time. Furthermore, these factors alone may not suffice in determining an appropriate layout design, as it is typically dependent on the specific task and data being utilised [197].

The discipline of *Graph Visualisation*, which is a sub-field of Information Visualisation, is primarily concerned with the challenge of *scalability* [78]. There are three types of limitations in this discipline [133]: computational capacity, human perceptual and cognitive capacity, and display hardware capacity. One example of limited computational capacity occurs when computer memory is insufficient to accommodate large datasets. Perceptual and cognitive capacity stems from the restricted capacity of visual working memory for information storage. Insufficient screen resolution due to display technology restrictions poses challenges in conveying all desired information simultaneously.

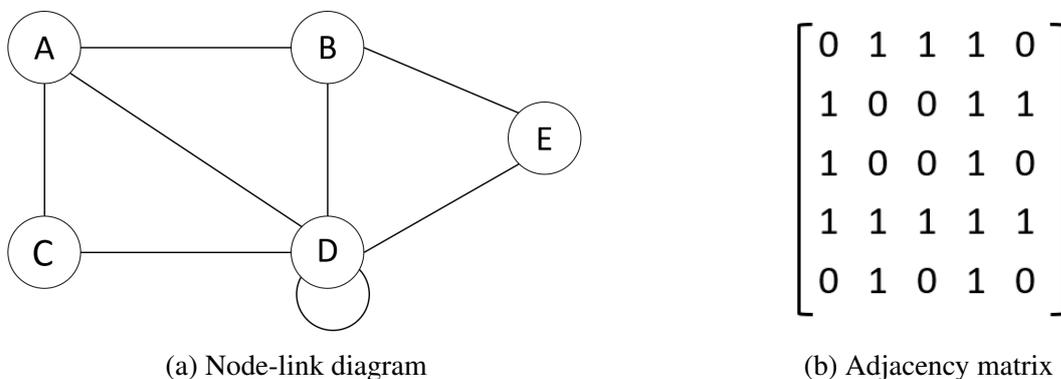


Figure 1.4 Two visual representations of the same network.

Graph drawing primarily emphasises the aesthetic criteria of node-link diagrams, but graph visualisation encompasses other visual metaphors that enhance its usefulness for the purpose of network exploration tasks. Our scope does not fall within the graph drawing research areas; this thesis focuses on graph visualisation approaches, which are more common in information visualisation.

The procedures and techniques employed to transform a graph into its corresponding geometric representation are commonly referred to as *graph drawings* [182]. The advancement of computer technology has facilitated the automated generation of graphs, hence fostering the progress of graph drawing. The *node-link diagram* is a widely used visual representation in

which vertices are displayed as points, and the connections between them are represented as lines [194], see Figure 1.4a.

In addition to utilising node-link diagrams for graph representation, an alternative approach involves implementing a more concise visualisation method known as the *adjacency matrix* \mathbf{A} , which can be binary. This matrix employs individual cells to denote the presence (represented by “1”) or absence (represented by “0”) of connections between nodes, with rows and columns serving as identifiers for the relevant nodes, see Figure 1.4b. The concept of this representation was initially proposed by Bertin [23] and subsequently gained popularity through the work of Becker et al. [21]. However, one limitation of this technique is its inability to describe graph structure, hence hindering the ability to perform path-related tasks [104, 137].

It bears noting that both node-link diagrams and adjacency matrices (or matrix representations) have distinct advantages and disadvantages. The selection of a visual representation relies on the specific characteristics of the network, the viewer, and the intended purpose of the visualisation. Careful selection of visual encoding that effectively communicates information while considering the cognitive limitations of human perception is crucial. For instance, a graph may employ misleading or ill-considered visual encoding techniques, such as assigning the same colour hue to nodes with different characteristics or utilising non-uniform sizes for nodes and edges to convey the importance of nodes and edges.

The motivation for this research is continued in the following section. It describes the primary issues that this thesis addresses.

1.3 Motivation and Research Questions

The volume of data generated on a daily basis around the globe is a significant obstacle in comprehending and deriving meaningful insights from it. Relational nature characterises numerous real-world data sets, including but not limited to transportation systems, computer security, and social networks. The data sets can be conventionally represented as graphs, whereby nodes symbolise entities, and edges symbolise the connections between them. On the other hand, alternative metaphors for graph representation, such as matrix representations, might be employed. The primary focus of this study revolves around node-link diagrams, as they offer a distinctive approach to comprehending relational data. Through the utilisation of human vision, it becomes possible to discern the underlying structure of the data, a task that would otherwise pose challenges in comprehension.

Visualisation serves as a fundamental method for conducting exploratory graph analysis. The process encompasses the creation of appropriate visual representations, such as adjacency matrices, node-link diagrams, or a combination thereof [77]. Additionally, it involves the effective arrangement of graph elements on the display and the efficient mapping of visual attributes [197]. We deliberately set the scope to node-link diagrams because they are the most common network visual representation for relational data in the field of graph visualisation [30].

Visualisations can be reproduced on physical media, such as printed on large-scale paper printouts, or shown in three-dimensional settings [132], (such as through virtual reality setups [42]), or projected onto displays, such as computers or laptops. Nevertheless, mapping

data in a three-dimensional environment can introduce complexities in perception. One such issue is that we can only project 3D scenes to a 2D display. Challenges such as occlusion and perspective arise, requiring the viewer to navigate several viewpoints (moving and rotating the object) in order to effectively comprehend the relationships among various data items [3]. It is more difficult to navigate in three-dimensional (3D) space because most input devices are designed with 2D in mind. Ware et al. [201] present a comprehensive examination of the 3D visualisation problems. They suggest that employing 2D or 2.5D solutions yields better results. Our research focuses on two-dimensional static representations, which can be effectively utilised in printed publications.

Uncertainty is omnipresent in our daily existence. It emerges whenever we are faced with choices or apprehensive about the future. Uncertainty visualisation is a subfield within the domain of information visualisation that primarily focuses on the presentation of visualised data in conjunction with additional information such as accuracy, error, and other characteristics related to the origin of the data that impact the comprehension and interpretation of the visualised information. [90].

On the other hand, graph visualisation primarily focuses on the representation of graphs using the main value of the input data, sometimes neglecting the consideration of its corresponding value uncertainty [197]. A considerable number of visualisations that we encounter fail to depict uncertainty [58]. One potential explanation for the absence of uncertainty representation in various visualisation scenarios, such as business reporting, media, and scientific contexts, could be attributed to the relative simplicity of presenting value in visualisations that do not incorporate uncertainty, as opposed to the considerably more complex task of demonstrating value in uncertainty visualisation [87]. Therefore, we require novel visualisation approaches to effectively communicate the uncertainty in the data to the general public.

The motivation for the research presented in this thesis is two-fold:

1. How can the most “important” edges of a large graph be identified so that it can be summarised and visualised efficiently? Given a large graph, how can the significance of a set of chosen subgraphs be measured? The first objective of this research is to identify short summaries for large graphs, with the aim of enhancing comprehension regarding the characteristics of the most “important” edges.
 - From a cognitive point of view, it is important to note the constraints imposed by human perception and its impact on the interpretation and comprehension of data [59]. For example, Miller’s ‘seven plus or minus two’ [130] is widely acknowledged in the field of cognitive psychology as a criterion that describes the capacity constraints of individuals’ working memory. Working memory can be considered as one of the ‘cognitive ceilings’ that may impact individuals’ capacity to reason about large networks. As a result, human working memory is limited, leading to challenges in processing and comprehending information when confronted with an excessive amount, resulting in cognitive overload and hindered sense-making abilities [120, 86].
2. How can we visualise graphs with additional data attached to the network edges and nodes in the form of a node-link diagram? The second objective of this research is to develop

novel visualisation techniques that represent both primary and secondary values, such as the uncertainty present in the data, while maintaining accurate communication.

- Research conducted in Psychology [96, 150, 158] has demonstrated that individuals exhibit a preference for receiving knowledge pertaining to uncertainty and make better decisions when such information is effectively presented. Moreover, appropriately conveying uncertainty establishes trust in various contexts. Despite the development of network visualisation tools and techniques, certain core challenges remain insufficiently addressed. Consequently, there is a growing need for practitioners and researchers in graph visualisation to explore novel concepts and develop effective approaches that can address the challenges posed by representing uncertainty. These techniques should facilitate users in effectively expressing uncertainty alongside primary value in a comprehensible manner.

In particular, this dissertation project aims at investigating the following research question:

- **RQ1:** Exploring a scalable approach for visualising complex graph-based data, including measures and variance of measures?
- **RQ2:** How to design a novel entropy-based bivariate representation of networks?
- **RQ3:** How do entropy-based representation networks compare in visual search performances to existing approaches?

1.4 Contributions

A novel approach for the visualisation of bivariate networks and a graph summarisation method is presented. The research adopts an interdisciplinary framework that integrates several methodologies from diverse fields. The methodologies employed in this study are tailored to the specific situation in which the research is conducted and the specific needs of the investigation. However, the essential concept underlying this node-link visual model is its general applicability to any bivariate network, irrespective of the specific domain of application.

To summarise, the main novel contributions of the thesis are:

- We present a novel node-link visual model — visual entropy (Vizent) graph — to effectively represent both primary and secondary values, such as uncertainty, on the edges simultaneously, see Chapter 4.
- We present the concept of the novel Vizent edge design and empirically demonstrate (in collaboration with Lucy McLaughlin) that different edge glyphs have a perceived order through pairwise testing, see Chapter 4.
- We perform two task-based usability studies to demonstrate the efficiency and effectiveness of our approach for visualising bivariate networks in the context of static node-link diagrams, see Chapter 5.

- We compare the Vizent design against three visual encodings selected from the literature on various graphs ranging in complexity from 5 to 25 edges for three different tasks, see Chapter 5.

1.5 Outline

The motivation behind this work, the goals of the thesis, and the criteria and methodology to achieve these goals have been described so far. This section provides an outline of the thesis alongside a brief overview of each chapter. The rest of this thesis is organised as follows:

Chapter 2 (Background) presents a contextual review of the literature with a discussion of the related work, beginning with the graph visualisation methods and issues of the graph summarisation methods, node-link visualisations, complex network data, uncertainty visualisation, and current research in graph summarisation.

Chapter 3 (Graph Summarisation Method) presents the proposed graph summarisation method to address information overload. It covers concepts of variance in edge weights, information theory and Shannon entropy, the entropy-based graph summarisation algorithm, testing with Facebook network data, limitations, and directions for future work.

Chapter 4 (Introduction Vizent Edges) focuses on how bivariate graphs can be visually represented within the node-link diagrams, specifically in the context of edge uncertainty. Furthermore, it introduces a novel visual encoding approach and presents the results of an empirical evaluation.

Chapter 5 (A novel node-link visual model – visual entropy (Vizent) graph) describes research comprising usability studies evaluating bivariate network visualisation techniques and discusses the results of two conclusive evaluations conducted to examine the novel node-link visual model presented in this thesis. It details the study design, tasks, data, results, and conclusions.

Chapter 6 (Conclusion) provides a summary of the key findings, including results of the quantitative and qualitative evaluations across the user studies. In addition, it summarises the thesis contributions and highlights promising directions for further research.

1.6 Related Publication

The majority of **Chapter 4** and **Chapter 5** are derived from the following published article, with each contribution being previously published in the *Visual Informatics* journal:

[1] Osman Akbulut, Lucy McLaughlin, Tong Xin, Matthew Forshaw, and Nicolas Steven Holliman, (2023). Visualizing ordered bivariate data on node-link diagrams. *Visual Informatics*, 7(3):22–36.

Abstract

Node-link visual representation is a widely used tool that allows decision-makers to see details about a network through the appropriate choice of visual metaphor. However, existing visu-

alisation methods are not always effective and efficient in representing bivariate graph-based data. This study proposes a novel node-link visual model — visual entropy (Vizent) graph — to effectively represent both primary and secondary values, such as uncertainty, on the edges simultaneously. We performed two user studies to demonstrate the efficiency and effectiveness of our approach in the context of static node-link diagrams. In the first experiment, we evaluated the performance of the Vizent design to determine if it performed equally well or better than existing alternatives in terms of response time and accuracy. Three static visual encodings that use two visual cues were selected from the literature for comparison: Width–Lightness, Saturation–Transparency, and Numerical values. We compared the Vizent design to the selected visual encodings on various graphs ranging in complexity from 5 to 25 edges for three different tasks. The participants achieved higher accuracy of their responses using Vizent and Numerical values; however, both Width–Lightness and Saturation–Transparency did not show equal performance for all tasks. Our results suggest that increasing graph size has no impact on Vizent in terms of response time and accuracy. The performance of the Vizent graph was then compared to the Numerical values visualization. The Wilcoxon signed-rank test revealed that mean response time in seconds was significantly less when the Vizent graphs were presented, while no significant difference in accuracy was found. The results from the experiments are encouraging and we believe justify using the Vizent graph as a good alternative to traditional methods for representing bivariate data in the context of node-link diagrams.

Chapter 2

Background

Contents

2.1	Human Perception	14
2.2	Graph	17
2.2.1	Main Graph Definitions	17
2.2.2	What is a Graph?	19
2.2.3	Graph Visualisation	22
2.2.4	Graph Layout	28
2.3	Visualisation of multivariate network	32
2.4	Uncertainty Visualisation	39
2.4.1	Uncertainty	39
2.4.2	Uncertainty Visualisation	40
2.4.3	Uncertainty Visualisation in Graphs	43
2.4.4	Summary	45
2.5	Graph Summarisation	46
2.5.1	Summary	50

The objective of this thesis is to describe a novel graph visualisation approach to representing bivariate data through a node-link diagram and suggest a summarisation algorithm for large networks. First, we start summarising the core concepts of cognitive and perceptual principles and demonstrate how these principles were utilised for supporting effective graph visualisation in Section 2.1. After, we introduce key graph terminology and properties in Section 2.2.1 and Section 2.2.2. Next, we briefly discuss related work in the context of graph visualisations and graph layouts in Section 2.2.3 and Section 2.2.4, followed by an analysis of node-link visualisations more closely aligned with our research objectives in Section 2.3. Furthermore, given that the design of Vizen Edge was specifically developed to depict additional data on edges, such as uncertainty present in the data, we present a concise overview of the term uncertainty and uncertainty visualisations in Section 2.4. Moreover, our other objective is simplifying complex relational data; thus, we gather a comprehensive survey of graph summarisation techniques in Section 2.5.

2.1 Human Perception

The human visual system encompasses the eyes, optic nerves, and specialised visual cortex within the brain. This intricate system processes and interprets visual stimuli, enabling perception and comprehension of our surroundings. *Information visualisation* leverages human visual-cognitive capacities for data analysis, exploiting its high-bandwidth “channel” communicating around 8.75 megabits per second of data to the brain [109]. With millions of photoreceptors and rapid parallel processing capacities, human vision can recognise patterns and scenes instantly [201]. This impressive visual bandwidth facilitates efficient data transfer from digital resources into the human mind. Nevertheless, a more significant advantage beyond mere data transmission is the human capacity for visual reasoning and extracting higher-level insights from data [33]. This ability allows individuals to construct mental models of the real phenomena embedded within datasets. According to Ware [201], the efficacy of visual perception in recognising shapes and features can be primarily elucidated through two psychological theories: *Pre-attentive processing theory* [188] and *Gestalt theory* [110].

The application of *pre-attentive processing theory* at a lower level pertains to the crucial matter of distinguishing one data object from another. Ware [201] presents an extensive compilation of pre-attentive properties (visual encodings), including line orientation, colour hue, curvature, basic shape, and size, that are pre-attentive processed, see Figure 2.1. Certain shapes and colours, for instance, appear to pop out against their surroundings; see Figure 2.2. The significance of such processes becomes highly pronounced in situations where the primary objective is visual search.

Gestalt theory, operating at a higher cognitive level, delineates the fundamental principles of the brain that underlie the understanding of visual images. [201]. The process of visual perception is widely recognised as a multifaceted phenomenon in which individuals have a tendency to perceive basic geometric forms. This implies that the structure constituting a visual presentation holds greater significance than the individual components, a concept encapsulated by the phrase "the whole is greater than the sum of its parts." *Gestalt theory* proposes a set of

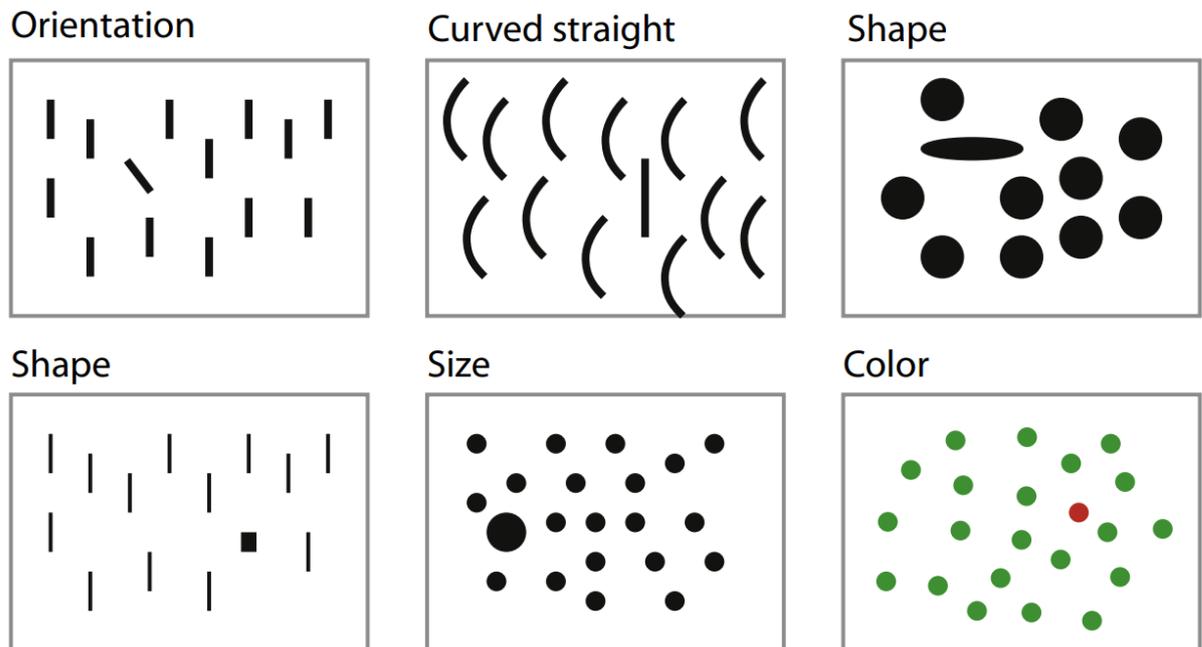
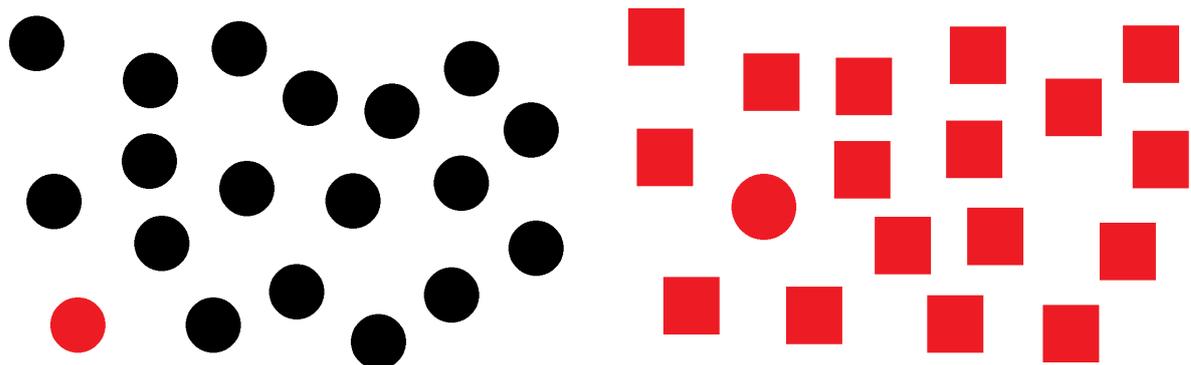


Figure 2.1 In the study by Ware [201] provided a comprehensive list of pre-attentive examples with a broad explanation, yet some of the pre-attentive features were given.



(a) The red circle "pops out" from a set of black circles.

(b) The red circle also pops out of a set of red squares.

Figure 2.2 Two examples of pre-attentive processing: spotting a red object from a set of black ones and spotting one circle from a set of red squares.

principles (or laws) that can enhance the intuitive comprehension and analytical reasoning of visual representations.

Gestalt principles encompass a set of guidelines governing the perceptual organisation of scenes. The concept of their introduction originated in the domain of philosophy and psychology during the 19th century. Subsequently, these principles were employed to establish fundamental principles of human perception in the early 20th century. The principles of Gestalt, derived from the German word for "form", encompass several key concepts. These principles (see Figure 2.3) include [108]:

- **Proximity**, which pertains to the grouping of closely positioned objects.
- **Similarity**, which pertains to the grouping of objects with similar shapes or colours.
- **Continuation**, which refers to the grouping of objects that create continuous patterns.
- **Symmetry**, which entails the grouping of objects that form symmetrical patterns.

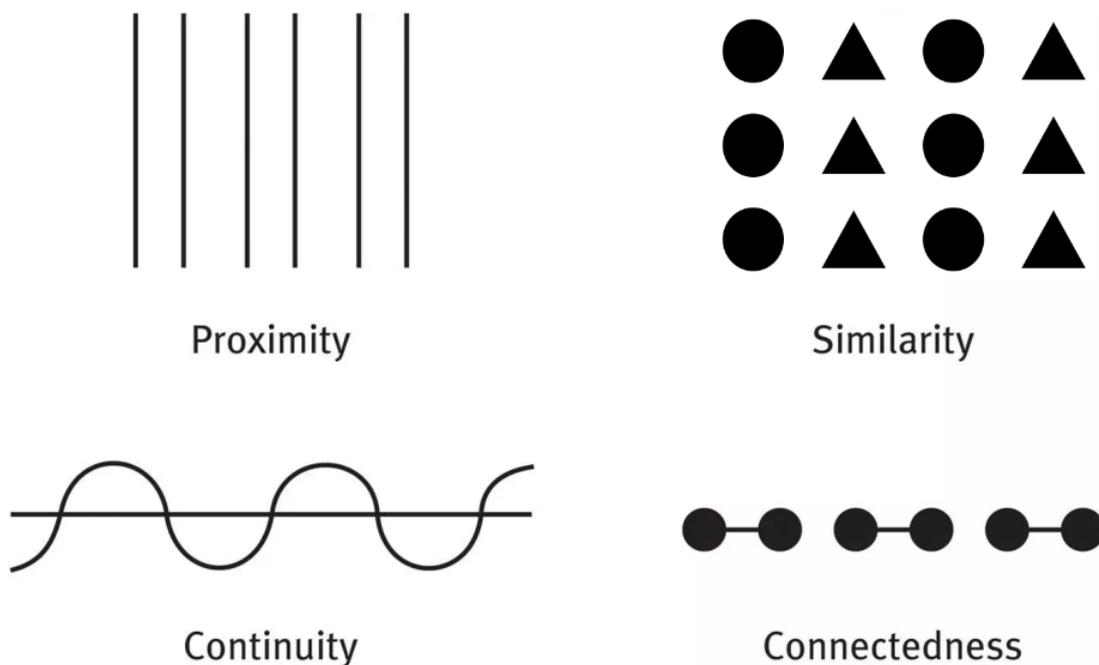


Figure 2.3 The figure represents a subset of Gestalt Laws (or Principles) [110].

The principles of Gestalt have found widespread application in the fields of user interface design, graphic design, and information visualisation. These principles provide a valuable framework for improving visual representations' effectiveness and aesthetic appeal in different domains. According to House et al. [85], "perception provides a sensible order to what we see whereas aesthetics govern our receptiveness to our perceptions." Ware et al. [202] conducted experimental studies that the utilisation of pre-attentive principles for encoding graph properties can facilitate rapid differentiation of these aspects in node-link diagrams.

In brief, while pre-attentive theory focuses on fundamental visual discrimination, Gestalt theory examines complex cognitive processes involved in conceptualising and integrating visual

information at a systemic level. These two serve as complementary lenses for enhancing the innate visual intelligence of humans, operating at distinct levels of abstraction.

Such perceptual organisation affects performance at visual search and other visual tasks. Comprehending these perceptual theories has been instrumental in shaping the Vizent approach. This study enhances the clarity and perceptibility of the primary value and its uncertainty in graph visualisations by implementing principles from Gestalt theory and pre-attentive processing. The development of the Vizent approach specifically targets the rapid recognition of data patterns within graphs, offering a significant improvement over traditional visualisation techniques that may overlook such perceptual optimisation.

2.2 Graph

2.2.1 Main Graph Definitions

“Your first discipline is your vocabulary;”
— Robert Frost

The term "vertices" is commonly referred to as "nodes", and "edges" are commonly referred to as "links" in academic literature. In this thesis, the interchangeability of these two terms is seen. A collection of fundamental terminologies pertaining to graphs was introduced as follows:

- A node j is referred to as a neighbour of node i if, and only if, there exists a connection between node i and node j .
- **Degree:** The number of edges that are linked to a certain node. The degree of node i is commonly denoted as $\text{deg}(i)$.
- **Walk:** A compilation of edges that are interconnected in a sequential manner to establish an uninterrupted route within a network.
- **Path:** A traversal that does not pass through any node (and, consequently, edge) more than once.
- **Cycle:** A walk that begins and finishes at the same node, without traversing any node more than once throughout its traversal.
- A **subgraph** is a subset of a larger graph that contains a subset of the nodes and edges of the original graph.
- A **connected graph** is a graph in which there is a path consisting of distinct edges that connect every pair of nodes.
- A **complete graph** that exhibits connectivity between every pair of nodes. In other words, a complete graph is a graph in which every node is connected to every other node by a single edge. Thus, it can be concluded that all complete graphs are connected; however, not all connected graphs are complete, see Figure 2.4.

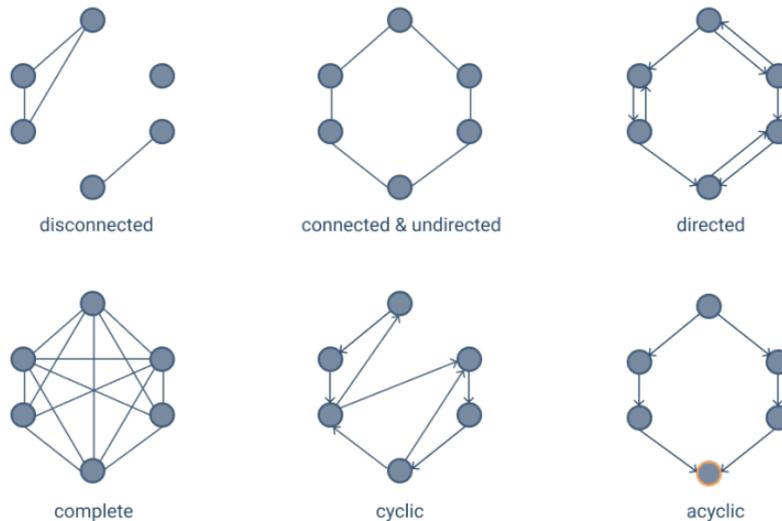


Figure 2.4 An example of different types of graphs.

- A **regular graph** is a type of graph where each node has the same number of adjacent nodes or the same degree.
- **Bipartite (n-partite) graph** is a graph wherein its nodes can be partitioned into two (or n) distinct groups, such that no edge exists between nodes within the same group.
- A **planar graph** is a type of graph that can be embedded in a two-dimensional plane without any of its edges crossing each other.
- An **undirected edge** refers to a connection between two nodes in a graph that does not have a specific direction or orientation. An **undirected graph** refers to a graphical representation consisting of undirected edges. The symmetry property holds for the adjacency matrix of an undirected graph.
- A **directed edge** consists of an ordered pair of nodes representing a relationship between two nodes. A graph consisting of directed edges is commonly called a **directed graph**. The adjacency matrix of a directed graph typically exhibits asymmetry.
- An **unweighted edge** refers to a type of edge in a graph that does not have an associated weight value. In a network with unweighted edges, the relationship between a pair of nodes can be categorised into two distinct possibilities: the presence of an edge connecting them or the absence of such an edge. The adjacency matrix of this network consists solely of binary values, specifically 0's and 1's.
- A **weighted edge** refers to an edge in a graph that has an associated numerical value or weight. This weight generally represents a certain attribute, which is a non-negative real number.

- **Multiple edges** are defined as connections between two nodes that occur more than once.
- A **self-loop** refers to an edge that connects a node to itself. A graph with an edge that originates and terminates at the same node is called a **Pseudograph**.
- A **simple graph** is a graph that lacks directed, weighted, multiple, or self-looping edges. Traditional graph theory is predominantly concerned with simple graphs.
- A **multigraph** is a graph with multiple edges. In the field of graph theory, it is possible for multigraphs to include self-loops. Multigraphs can be undirected or directed.
- A **dense** graph can be defined as a graph where the number of edges increases by the square of the number of nodes, while a **sparse** graph has a lower rate of edge growth.

Graph theory is a field that involves a range of network properties, such as density, diameter, and clustering coefficient. However, graph theory is not the subject of this thesis.

2.2.2 What is a Graph?

Graphs are generally suitable for representing entities where a network structure is to be represented [30]. Graphs serve as a unifying theme within the field of computer science since they provide an abstract depiction of the structural arrangement of various systems, including transportation systems, human interactions, and telecommunication networks [174]. In order to facilitate the understanding of the graph-based data structure, it is necessary to know its basic structure.

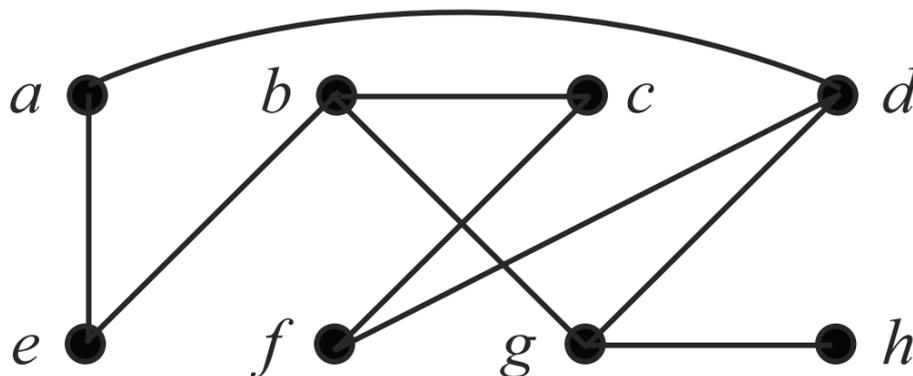


Figure 2.5 A visual representation of the graph G .

Formally, a plain (simple) **graph** ' G ' := (V, E) or network is an abstract data type consisting of a finite set of vertices (or nodes) V and a set of edges (or links) $E \subseteq \{(u, v) | u, v \in V; u \neq v\}$. The latter $e_{u,v} = (u, v) \in E$ connects a pair of nodes u, v , denoting that these nodes are directly related in a meaningful manner. If E is not a multiple set and does not contain self-loops, then the graph is *simple*, see Figure 2.5. For example, an edge (u, v) is a self-loop if $u = v$; see green edges in Figure 2.6. If node pairs can have multiple links between them, then the graph is *multigraph*, see Figure 2.6.

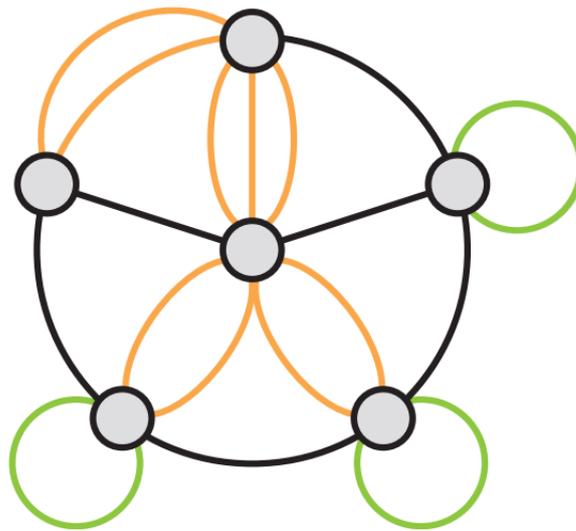


Figure 2.6 An example of a multigraph with multiple orange-coloured edges. [26]

As seen in Figure 2.5, The set V consists of the elements $\{a, b, c, d, e, f, g, h\}$, whereas the set E is comprised of the following pairs: $\{\{a, d\}, \{a, e\}, \{b, c\}, \{b, e\}, \{b, g\}, \{c, f\}, \{d, f\}, \{d, g\}, \{g, h\}\}$. The graph G is formed by the combination of vertices V and edge E . The number of nodes and links present within a graph can be counted. As an illustration, the graph depicted in Figure 2.5 consists of a total of eight nodes and nine edges. The graph *size* is equivalent to the total number of nodes within the graph.

Graphs are commonly categorised into two main types: *directed* graphs and *undirected* graphs. If each edge is represented by an unordered (ordered) pair of vertices, then the graph is undirected (directed). It should be noted that while the ordered pair (v_1, v_2) is *distinct* from the pair of (v_2, v_1) in a directed graph, they are inferred as the same relationship in an undirected graph. In an undirected graph, (v_1, v_2) and (v_2, v_1) are the same edge, and only one pair is enough to represent. A graph with both directed and undirected edges is referred to as a *mixed* graph.

In the context of an undirected network, relationships are regarded as bidirectional, such as friendships. In the context of a directed graph, it is important to note that relationships between nodes possess a distinct directionality. Relationships directed towards a specific node are commonly known as *in-links*, whereas relationships originating from a node are referred to as *out-links*. The inclusion of direction introduces an additional dimension of information. Relationships of the same nature but with opposite directions possess distinct semantic connotations, signifying either a state of dependency or a directional movement.

Graphs can also exhibit bipartite characteristics, meaning they consist of two separate sets of nodes, U and V , representing various entities. In such networks, each relation that connects a node from set U to a node from set V is called *bipartite graph*, as seen in Figure 2.7.

Cycles refer to the paths traversed inside a network of interconnected nodes and relationships, wherein the journey commences and concludes at the identical node. An acyclic graph is characterised by the absence of cycles. In classic graph theory, a *tree* (a special case of a graph) is defined as a type of linked undirected graph that does not contain any cycles. In the field of computer science, it is worth noting that trees can also possess a directed nature. A more

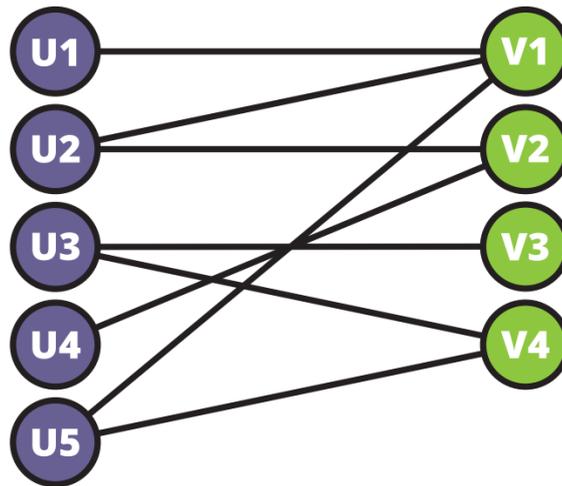


Figure 2.7 Illustrating a bipartite graph that has two node types. [26]

comprehensive characterization may be posited as a network in which a singular path exclusively facilitates the connection between any two nodes.

A tree T is considered rooted when a specific vertex r is designated as the *root node*, see Fig 2.8. This can be denoted as ' $T := (V, E, r)$ '. These trees are commonly seen as **hierarchies**, in which the level of a vertex in the hierarchy is determined by the length of the path to the root. In a formal context, it is important to note that a **hierarchy** is defined as a directed acyclic graph. Consequently, inside a formal hierarchy, it is possible for a node to possess many pathways leading to the root node. Tree visualisation, also known as hierarchy visualisation, is a subfield of information visualisation that focuses on the graphical depiction of connected, acyclic graphs — trees.

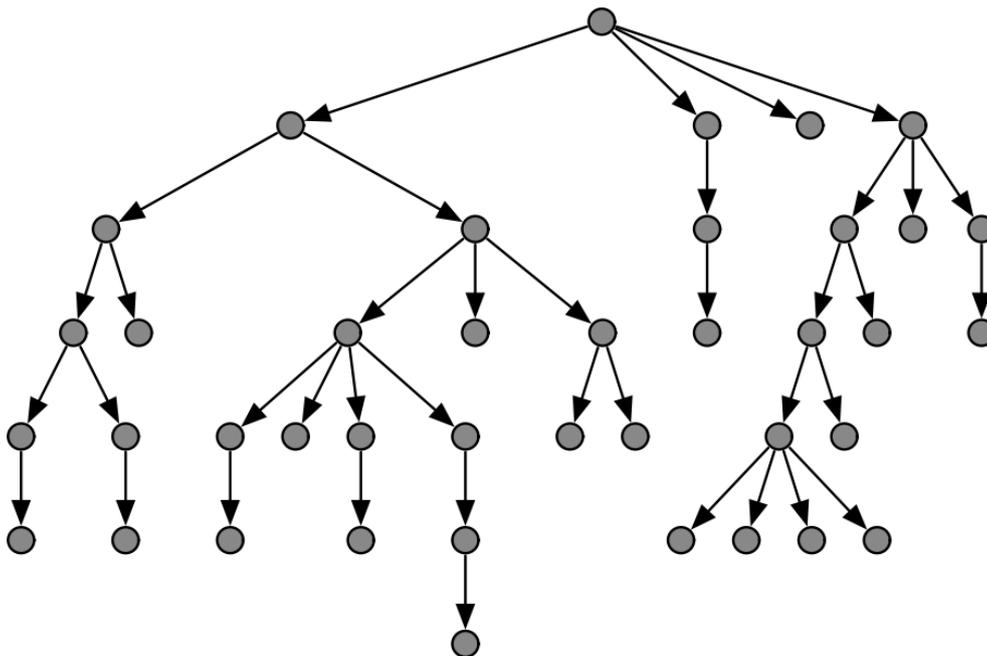


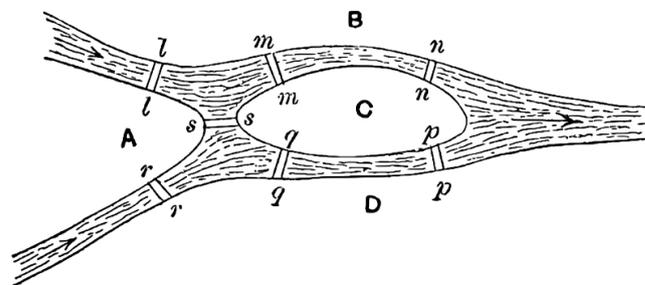
Figure 2.8 Illustrating the representation of a rooted tree utilising the *TreeLayout* algorithm [37].

2.2.3 Graph Visualisation

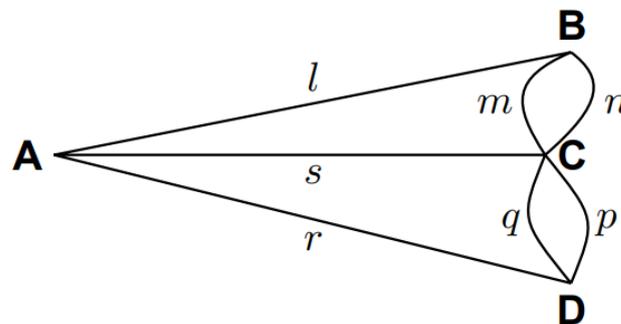
Graphs have been commonly employed as a means of representing relationships between entities for several purposes. Some examples of graphs include telecommunication networks, social network analysis, traffic networks, computer networks, co-authorship networks [115], protein-protein interaction networks [92], and functional or structural brain connectomes, among various others. Therefore, graph visualisation has been thoroughly researched within the literature and remains an active research area. This section surveys relevant efforts in graph visualisation.

Historical Development

Node-link diagrams have a historical origin in the 18th century, specifically in the context of the "seven bridges of Königsberg" problem. This problem was initially formulated by Leonhard Euler, who employed nodes to represent distinct regions inside the city and links to represent the bridges connecting these regions. In 1736, Euler released his famous paper on the city of Königsberg. He addressed the path-tracing problem by employing a graph structure consisting of interconnected nodes and edges. Euler's primary objective was to address the inquiry of whether it is feasible to get from one island to another while visiting each bridge just once. This milestone signifies the shift from early graph representations to contemporary graph visualisation techniques. Nevertheless, the graph proposed by Euler exhibits characteristics that resemble a visual representation rather than a formal mathematical network (see Figure 2.9a).



(a) Euler's drawing [54]



(b) Ball's abstract drawing [14]

Figure 2.9 Node-link graph drawings of Königsberg problem

The initial manifestation of an abstract graph drawing may be traced back to Ball's publication on mathematical recreations [14] when he presented a revised depiction of the Königsberg problem through the use of node-link diagrams (see Figure 2.9b). This significant development occurred in the year 1892, more than a century and a half after the original challenge emerged.

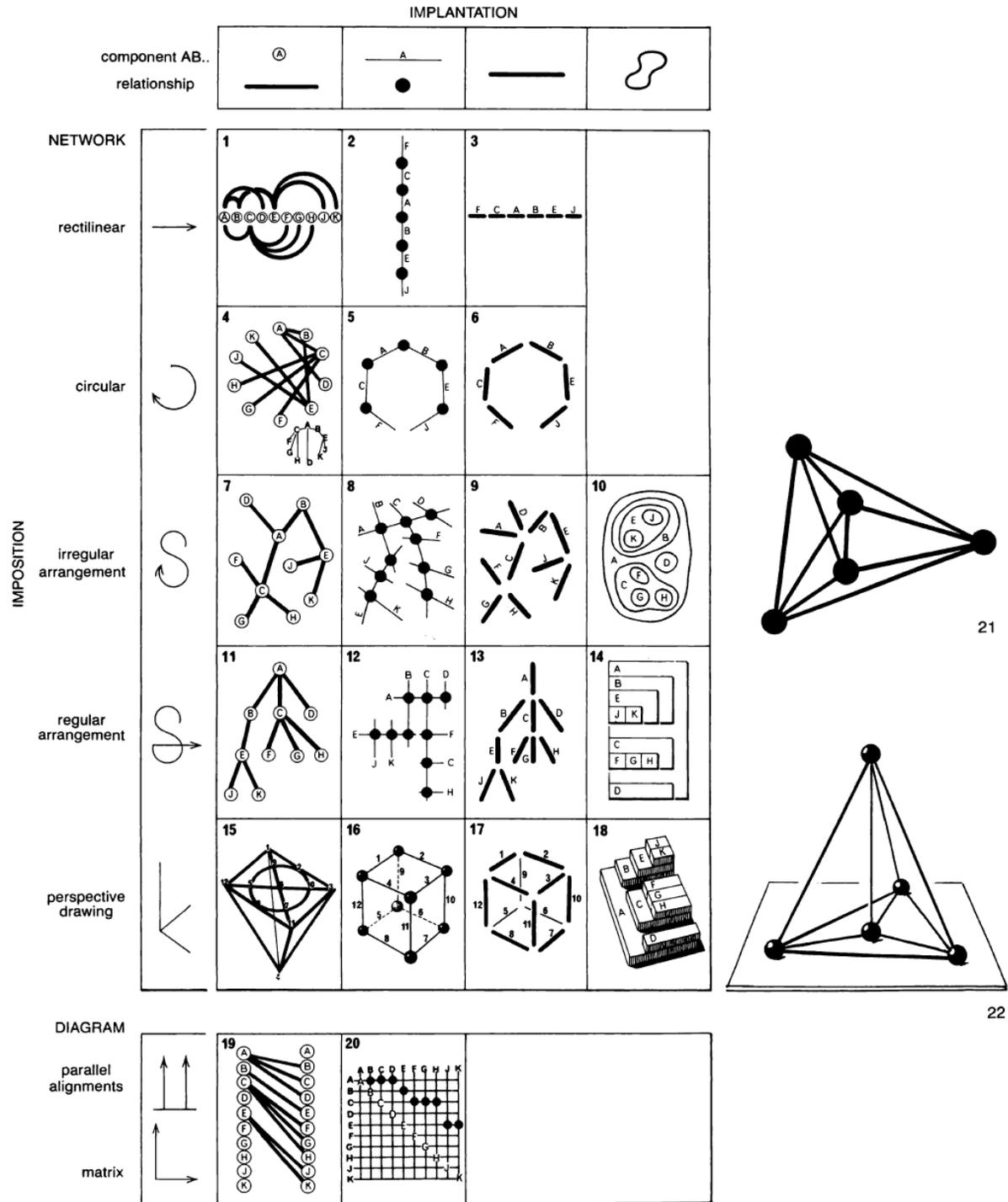


Figure 2.10 Bertin [23] proposed various visual encodings for representing graphs.

Bertin [23] categorises a total of twenty-two visual encodings, classifying them into two distinct categories: **networks** and **diagrams**, see Figure 2.10 above. Bertin's conceptualisation of *networks* refers to visual depictions wherein each vertex and edge is singularly portrayed, such as in traditionally used **node-link diagrams**. In contrast, Bertin provides a definition of a

diagram as a visual depiction wherein every node is depicted twice, serving as the two endpoints of an edge. Examples of such diagrams include matrices and parallel alignments, which are also denoted as bipartite layouts. An exemplification of a bipartite layout can be seen in Figure 2.11.

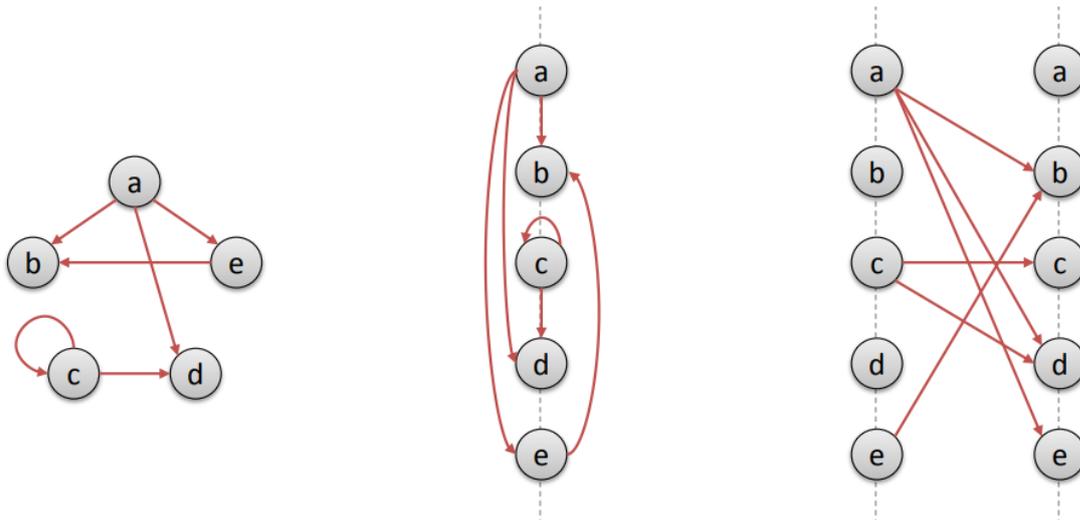


Figure 2.11 The diagram illustrates the procedure for converting a pseudograph into a graph employing a bipartite layout, which involves duplicating the vertices and arranging them equidistantly along one axis of two parallel vertical axes. The edges are then represented as straight links connecting the vertices on the two axes. [20]

Bertin’s classification of networks differentiates between planar variables and retinal variables. Planar variables encompass the “implantations” and “impositions” (see Figure 2.10). The term “implantation” (which means *instantiation*) refers to classifying geometrical primitives (or symbols), namely points, lines, and areas. On the other hand, “imposition” refer to the spatial arrangements of these geometrical primitives, such as different types of layouts.

Based on Bertin’s analysis, there are four distinct graphical representations for graphs. The first type involves the utilisation of points to depict vertices and lines to symbolise edges. The second type reverses this approach, where points represent edges and lines represent vertices. The third form exclusively employs lines to represent edges without explicitly visualising the vertices. Lastly, the fourth alternative utilises areas to represent both edges and vertices simultaneously. Bertin’s retinal variables encompass six basic additional characteristics (visual channels) of “implantations”: size, colour hue, value, orientation, texture, and shape.

Regarding the use of the planar dimensions, Bertin classifies them as rectilinear, circular, irregular, regular, and perspective drawings. It is imperative to recognise that not all graph visualisations are appropriate for every graph category. Specific visualisations, such as area inclusions (also known as enclosure or containment) or space-filling approaches, have been developed for particular graph types, especially for the purpose of visualising hierarchies and other tree structures.

Space-Filling Approaches

In the literature, various different graph representation metaphors for space-filling layouts of trees have been employed [201], such as *Treemaps* [195], *Radial trees*, *Icicle plots*, *Sunburst* [179], *Circular treemaps* and *BeamTree* [193] methods, see Figure 2.12.

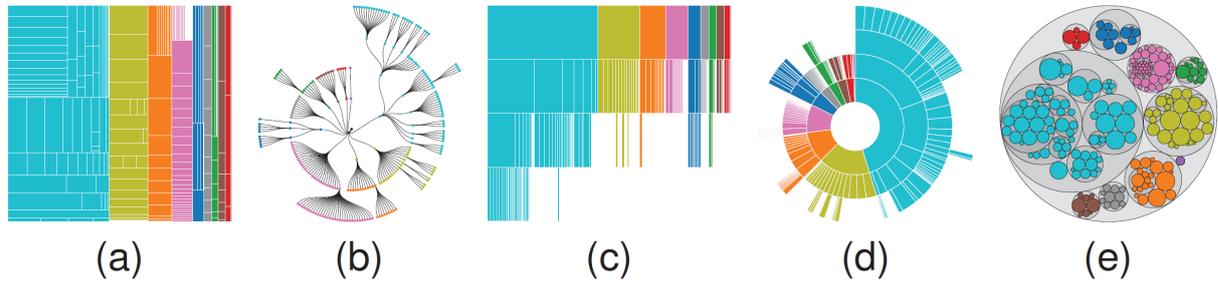


Figure 2.12 Traditional tree-based visualisation techniques for hierarchical data in which colour encodes to univariate categorical data. Treemaps (a), radial trees (b), icicle plots (c), sunburst charts (d), and circular treemaps (e). [213]

One of the most notable instances is the utilisation of *Treemaps*, a hierarchical visualisation technique that was introduced by Johnson and Shneiderman [94]. Treemaps involve using rectangular shapes that are recursively subdivided inside a display area based on the underlying hierarchy, as shown in Figure 2.13. Furthermore, there are different versions of Treemap algorithms that can either implicitly encode only inner nodes through node arrangement or generate a boundary outline for inner nodes [135]. One possible method of encoding a secondary attribute is using colour hue or value/saturation. However, it is important to note that alternative encodings, such as glyphs [189] or approximate positioning [175], can also be employed. However, we will not provide further explanations on space-filling alternatives, as this thesis primarily focuses on the node-link metaphor.

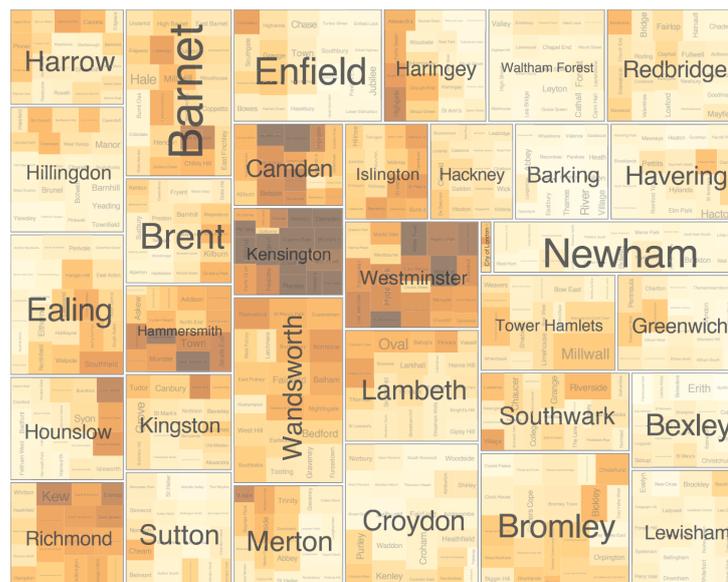


Figure 2.13 Illustration of a Treemap layout that depicts housing data pertaining to several boroughs of London. Attributes are represented by the encoding of size, colour, labels, and an approximation of position. [175].

Node-Link and Matrix Representation

The domain of graph visualisation encompasses a wide range of surveys conducted on various forms of graph data, indicating its extensive scope. Von Landesberger et al. [197] present a taxonomy of visual representations based on their dependence and structural characteristics. Additionally, the authors conducted surveys that examine diverse methodologies employed in the visual analysis of large graphs, focusing on their ability to manage scalability challenges. The study encompasses various aspects, including visual graph representations (**static** and **dynamic**), user interaction (such as panning and zooming), visual graph analysis, and future challenges (such as edge visualisation).

The categorisation of static graph visualisations commonly involves distinguishing between node-link and matrix representations, see Figure 2.14. Node-link diagram is recognised as the most prominent and extensively employed approach for graph visualisation [191]. They are intuitive and effective in facilitating the perception of relationships between objects and solving path-related tasks. This is mostly due to their ability to leverage the concepts of closure and good continuation, as outlined in Gestalt psychology [110]. Therefore, node-link diagrams are mostly suitable presentation of relational data for human perception to enhance understanding of the features and relationships of graphs [19].

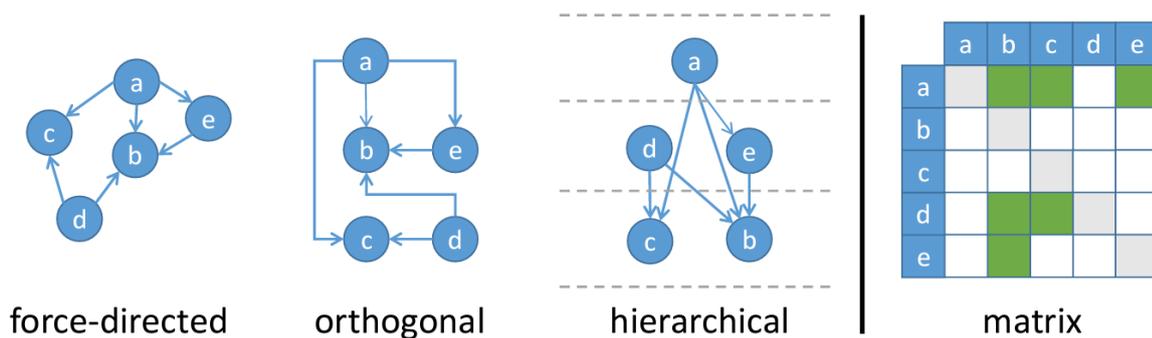


Figure 2.14 Example of the different visual representations of static graphs. All representations show the same dataset [19]. Left three images: node-link diagrams. Right: Adjacency matrices (matrix).

As an alternative to node-link diagrams, matrix-based representations are also utilised [68, 136]. These representations may address certain challenges encountered in node-link diagrams, particularly when dealing with large and complex networks. These include node occlusion and edge crossings [173]. Conversely, the perception of spatial characteristics may be compromised, resulting in increased difficulty in tasks such as locating nodes along a path and discerning clusters. In addition, novice users may encounter challenges when understanding matrices [164]. Moreover, the readability of matrices deteriorates significantly when the number of nodes increases substantially, resulting in excessively small rows and columns.

Numerous studies have conducted comparisons of node-link diagrams utilising certain graph layouts and adjacency matrices visualisations across a diverse range of tasks [68, 67, 103, 137]. The findings demonstrate that the effectiveness of the visualisation is significantly influenced by the characteristics of the provided dataset and the assigned tasks. A study conducted by Ghoniem

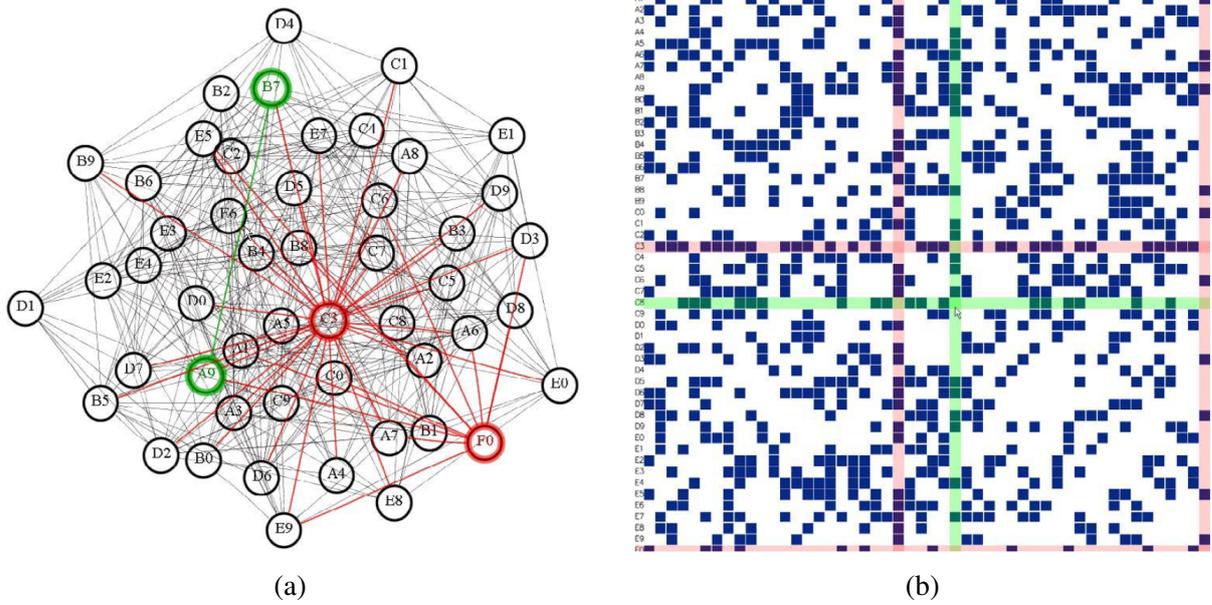


Figure 2.15 Evaluation of node-link diagrams and adjacency matrix visualisations [67].

et al. [68] demonstrated that the effectiveness of the two visualisations in facilitating particular tasks is contingent upon the size and density of the graph.

Hybrid Graph visualisation

The literature also contains methodologies that combine the strengths of the two previous approaches. [76, 77]. One study [77], for instance, incorporates both representations in a single view, wherein node-link diagrams are employed to depict the overall graph structure of the network, while adjacency matrices are utilised to visualise the communities inside the network, see Figure 2.16. Additionally, adjacency lists can be used for graph visualisation, although they are more commonly used as a space-efficient data structure for handling graph data [79].

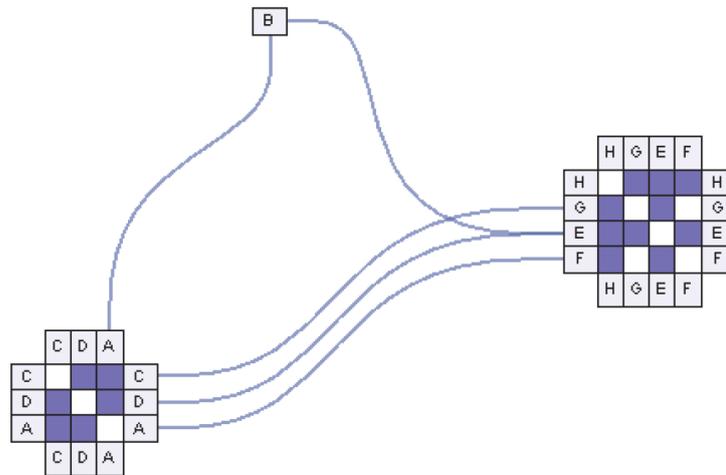


Figure 2.16 Hybrid graph visualisation combining node-link diagrams with adjacency matrices [77].

2.2.4 Graph Layout

As presented by Bertin's work [23], one of the conventional methods of representing objects involves using dots to symbolise each object, with lines drawn between objects to indicate connections. *Graph layout* refers to the process of determining the spatial arrangement of nodes to provide an *aesthetically* pleasing representation and enhance the comprehension of the underlying information for end-users. Graph layout methods can provide 2D or 3D placements to the nodes within a particular graph [117]. The selection and implementation of specific graph layout algorithms can provide valuable insights into the topological characteristics of a graph. Node-link diagrams possess a noteworthy benefit in terms of conserving the local characteristics of a network, hence facilitating the identification of its neighbours to a given node and enabling the tracking of paths within the network. However, if these algorithms are not chosen or implemented appropriately, they may obscure the inherent nature of the network's structure.

Battista et al. [17] and Eades et al. [51] examined different methodologies employed in graph drawing and provided aesthetic principles for making drawings better. Additionally, Herman et al. [78] conducted a comprehensive examination of graph visualisation and analysed it within the information visualisation framework. In addition to incorporating a broader range of visual representations than node-link diagrams, this study examined and compared conventional graph layouts, such as tree-based, hyperbolic-based layouts, and 3D.

Optimising display space utilisation is thoroughly explored within the *Graph Drawing* community. Considerable effort is devoted to the computation of two-dimensional layouts for node-link diagrams while taking *aesthetic criteria* into account to enhance the readability of these diagrams [197]. Common criteria for graph layout often involve ensuring non-overlapping nodes, minimising the number of edge crossings, maintaining uniform edge lengths, maximising symmetry and facilitating the easy recognition of graph substructures. Concerning the previously described *aesthetic* objectives and limitations, the objective of the graph drawing community is to identify algorithms that use the display area effectively and thus generate acceptable solutions.

Numerous layout methods have been presented to generate a visual arrangement of nodes and edges in a network within the Graph Drawing community, serving as important aspects in graph visualisations, such as radial, circular, and force-directed layouts [61]. These algorithms mostly rely on node-link diagrams. Utilising various layouts of the same dataset can significantly influence our perception of the relationships among data objects.

The random layout algorithm allocates a random x and y position to each node while ensuring these positions fall inside the specified display areas, see Figure 2.17 below. The speed of the generated layout is fast, and although the underlying network structure remains unidentifiable, the visual representation of the network's nodes and edges might provide a first assessment of its size and density. In graph user studies, random layouts are occasionally employed to establish a benchmark for evaluating the performance of more advanced layout algorithms. [30].

The circular layout algorithm uses a circular reference system to determine the positioning of nodes. All nodes are exclusively positioned on the circumference of the circle. The node sequence in a network can be determined by several criteria, such as alphabetical arrangement,

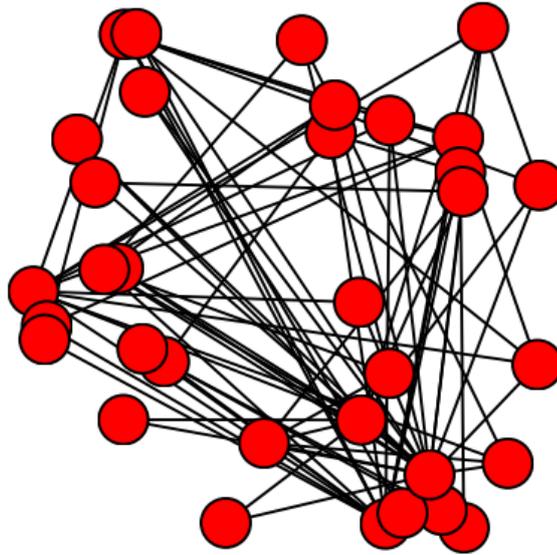


Figure 2.17 Random layout [160]

sorting based on node degree, node similarity, or network clustering. Circular layouts that have node sequences of significance can provide valuable information [26]. Employing a circular layout in which nodes are sequenced according to their relative number of connections (node degree) can be one example of this scenario. This arrangement aids in the identification and understanding of the overall interconnectedness present within a network.

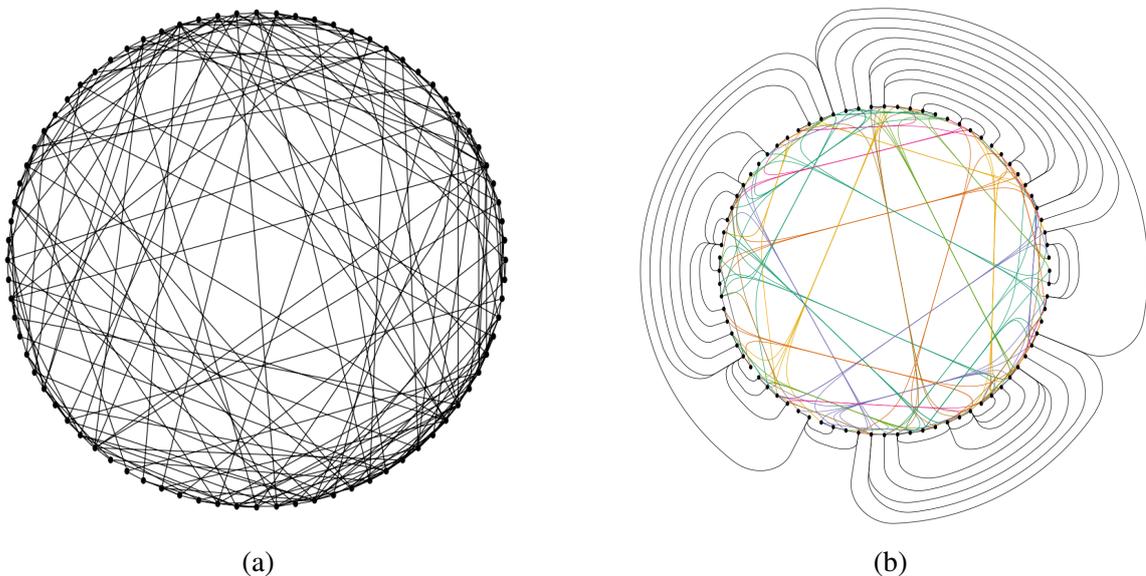


Figure 2.18 a) Circular layout employing edge-length minimising order. b) Circular layout employing exterior routing with edge clustering algorithm [65].

In their study, Gansner and Koren propose an *improved circular layout* algorithm that routes edges along either the outer or inner face of the circle. The edges that are routed inside the circle

are drawn using an edge bundling technique that aims to improve the utilisation of the available area, see Figure 2.18b.

Force-directed approaches have been the predominant focus in the literature on network layout [36], mostly due to their simplicity, ease of implementation and ability to generate reasonable link visibility [173]. This approach involves calculating attractive and repulsive forces between pairs of nodes in order to identify their optimal proximity in the resulting graph visualisation. These forces are incrementally exerted upon the nodes until a state of stability is achieved. These methods can be applied to various types of graphs without requiring prior knowledge about the underlying structure of the graphs. One advantage of this strategy lies in its simplicity, as a basic iteration can be effortlessly executed. Holten et al. [83] introduced a force-directed algorithm in which links are modelled as flexible springs that can attract each other while keeping node locations unchanged. As illustrated in Figure 2.19, the traditional node-link diagram employs a force-directed layout that incorporates size encoding to represent edge and node attributes.

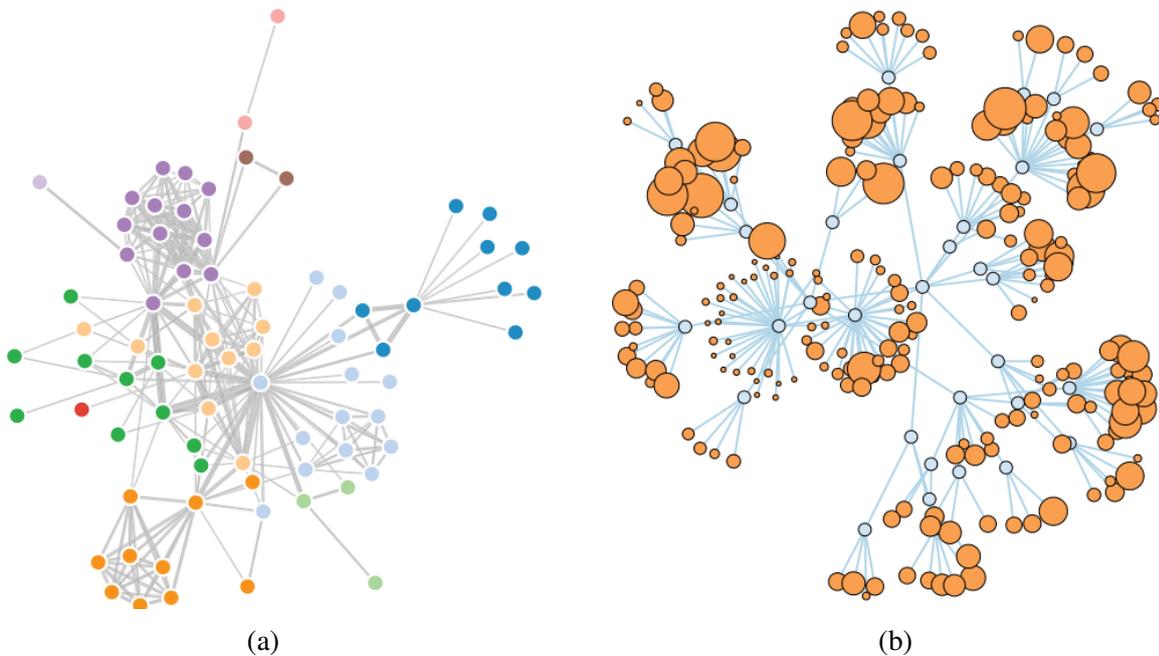


Figure 2.19 a) Illustration of a force-directed layout algorithm applied to a graph consisting of 75 nodes. The layout incorporates size encoding to represent edge properties. b) A larger network where the node size encodes node attributes.

Relational data sets that incorporate actual reference systems, such as geographic maps, can utilise network overlays that employ the geographic location of nodes to establish their spatial locations, see Figure 2.20. One prevalent method for representing geo-located graphs is through a map that exhibits a portion or the entirety of the earth, where nodes and edges are accurately positioned based on their geospatial locations [28].

Extensive research has been conducted on various graph layouts for a long time. Several earlier studies investigated how individuals visually interpret these visualisations and assessed which visualisation methods are more effective in supporting particular tasks and datasets within node-link diagrams. For example, a study conducted by Purchase et al. [146, 148] examined the

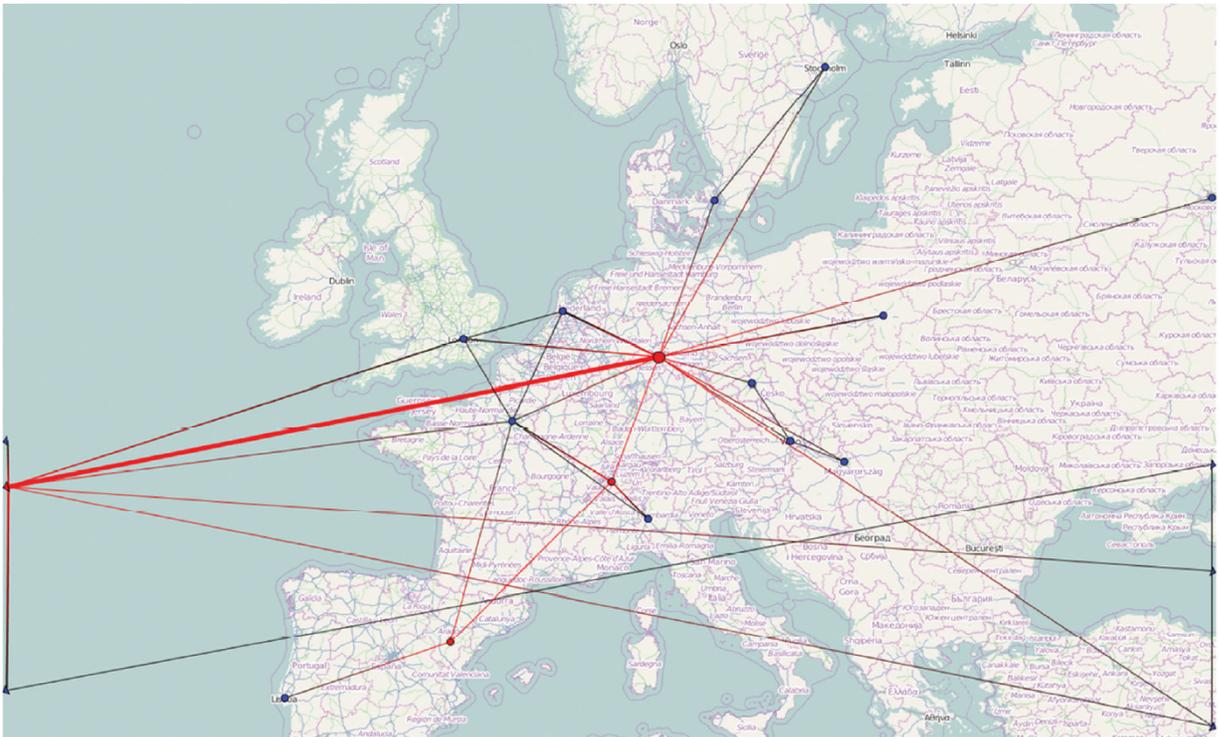


Figure 2.20 Illustration of a graph visualisation on a map with geographic location [28].

impact of different graph layouts on the readability of visualised data in the node-link diagrams. Ware et al. [203] investigated the impacts of various layout characteristics on response time for a shortest-path finding task involving node-link diagrams.

Selecting a graph layout method is a critical step in implementing any network visualisation environment. Traditionally, an adequate layout technique would effectively depict the underlying graph structure and mitigate the scalability issue, which is a persistent concern faced by the broader information visualisation community. The undertaking of building an effective graph drawing algorithm is frequently complex and requires a significant investment of time. Consequently, there exists a variety of open-source packages, such as Cytoscape [168], Gephi [15], and Graphviz [53], among others, that address the matter of graph layout concerns. In the majority of instances, while managing small-scale networks, the available resources are typically considered satisfactory.

Conventional methodologies prioritise identifying graph structures, specifically topological characteristics while facilitating interactive exploration [7]. Moreover, numerous graph analysis tools heavily depend on statistical techniques to investigate various graph structure properties, such as distances and node degrees. Nevertheless, the provided tools have limited functionality in expressing node and edge attributes.

Considerable efforts have been made to compute optimal two-dimensional layouts for node-link diagrams. These layouts aim to communicate the graph topology effectively while incorporating aesthetic criteria to enhance legibility [17]. The domain of graph layout research is vast, and conducting comprehensive research on all proposed methodologies is beyond the scope of this dissertation. Our research does not focus on graph drawing algorithms. Instead, we concentrate on the visual design of node-link diagrams to assist in exploring the graph. A crucial

aspect to consider in designing node-link diagrams is effectively employing an appropriate visual mapping of nodes and edges to convey associated attributes.

2.3 Visualisation of multivariate network

Graphs generally include additional information. An attribute encompasses information pertaining to the *characteristics* or *properties* of a network data instance. In addition, attributes can be attached to nodes ($a_i \in A_{nodes}$) and edges ($a_j \in A_{edges}$) in order to indicate various metrics, such as their type, size, or other information relating to its application. Edges frequently signify a relationship's weight (such as its strength and significance) as well as its direction. Conversely, attributes can also be derived by computing topological properties, such as node degree and node centrality [135].

A multivariate network is formally represented as $\bar{G} := (\bar{N}, \bar{E}, \rho)$, wherein \bar{N} denotes a set of nodes, $\bar{E} = (E'_1, E'_2, \dots, E'_k) : E'_j$ denotes a set of all edges with j^{th} attribute value $A_j \in \{1, 2, \dots, k\}$, and ρ signifies a function that maps each node to its corresponding attribute vector $n \in \bar{N}$.

According to Kerren et al. [104], in the simplest form, the concept of multivariate networks encompasses an underlying graph \mathbf{G} , along with \mathbf{k} additional attributes that are associated with either nodes, edges, or both. Multivariate network visualisation involves enhancing traditional network visualisation methods by incorporating supplementary variables linked to the graph's nodes and/or edges. The most difficult aspect of visualising multivariate networks lies in the simultaneous representation of both the inherent network topology and its corresponding attributes [135].

Multivariate data is commonly observed in real-world graphs pertaining to nodes and links. Enhancing the representation of a system provided by a graph can be achieved through the incorporation of additional attributes to the nodes and edges. This augmentation aims to unveil additional properties inherent to the system under evaluation [30].

Datasets containing a single dimension are commonly referred to as **univariate**, while datasets having two dimensions are known as **bivariate**. On the other hand, the term **multivariate** is generally used to describe datasets encompassing three or more variables. In the context of node-link diagrams, the prevailing approach involves utilising a single data variable to visually encode graphic symbols, specifically nodes and edges. This type of symbol encoding is referred to as "univariate". However, many traditional graph visualisation techniques fail to account for the presence of additional attributes associated with each graph item, which are crucial for effective representation [152].

Bertin's seminal work proposed seven visual variables: location, size, colour hue, colour value, texture, orientation, and shape [23]. Mapping data to visual variables, such as size, and colour, allows end-users to discern unseen patterns and derive valuable insights from the data [98]. Numerous studies have explored visual channels, such as fuzziness and colour saturation, for depiction [98].

The visualisation of edge directionality is frequently represented using arrows positioned at the ends of the edges. However, alternative methods such as tapering or gradients can also be

employed. [82]. Text labels are commonly employed for visually representing nodes and edges attributes; nonetheless, visual channels, including colour hue, node size, varying edge thickness, and shape, are also utilised [104]. For example, Bach et al. [11] exclusively employed node size as a means to visually signify node degree in the exploration of network evolution.

Karim et al. [98] conducted research to examine the effects of colour coding in the context of node-link diagrams. The researchers examined the impact of four distinct colourmaps, both single- and multi-hue colourmaps, namely Blue, Viridis, RdYlBu and Jet, on the representation of quantitative node properties inside a node-link diagram. In the experiment, a node-link diagram with three emphasised nodes was provided to the participants. The circle-shaped node served as a reference point, while the two square-shaped nodes were those to be compared, see Figure 2.21. Their investigation revealed that Blue colour maps and Viridis demonstrated notably enhanced performance, yielding reduced error rates.

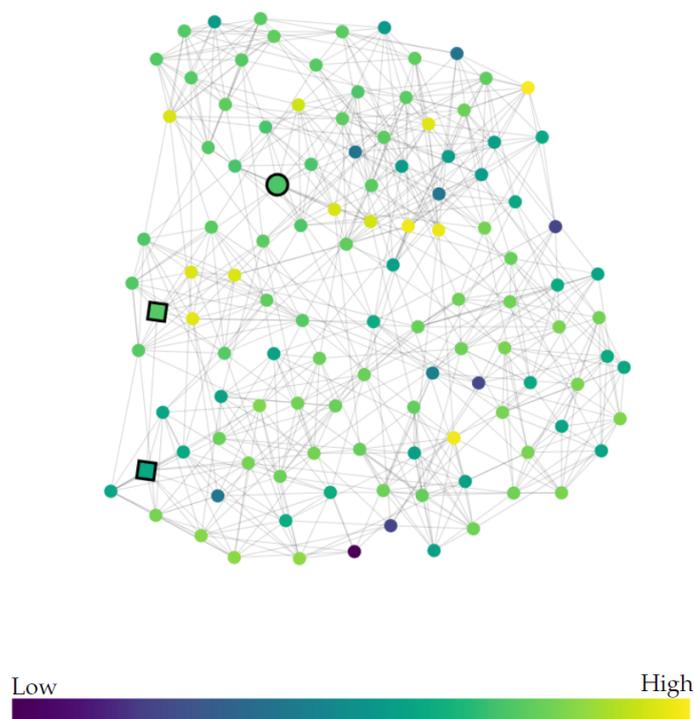


Figure 2.21 The evaluation of colour encoding on nodes was conducted in node-link diagrams [98].

In their work, Ghani and Elmqvist [66] examined the efficiency of different visual channels used for node mappings concerning revisitation tasks. The researchers conducted a comparative analysis of three distinct visual mapping techniques for representing nodes: size, colour hue, and a combination of size and colour, see Figure 2.22. The study results indicated that incorporating spatial location into node size and colour yields better outcomes than using either size or colour alone as node encoding techniques.

A further node-link diagram example encodes attributes on nodes and edges through colour and size visual channels [2]. Each node denotes a scientific article with colour indicating the publication year and size representing global citation counts. Edges represent bibliographic coupling relationships, with colour interpolating from the nodes' colour, as can be seen in Figure 2.23.

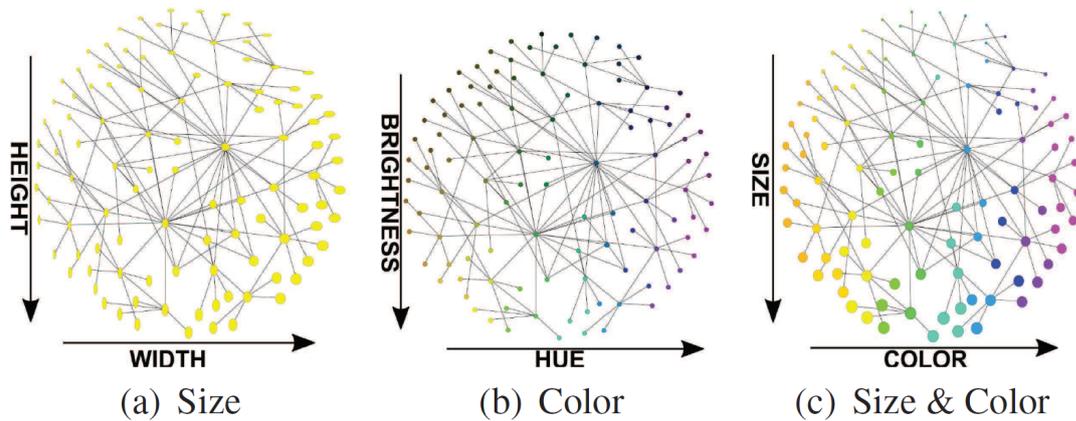


Figure 2.22 Three different visual mappings were evaluated on nodes in the context of node-link diagrams [66].

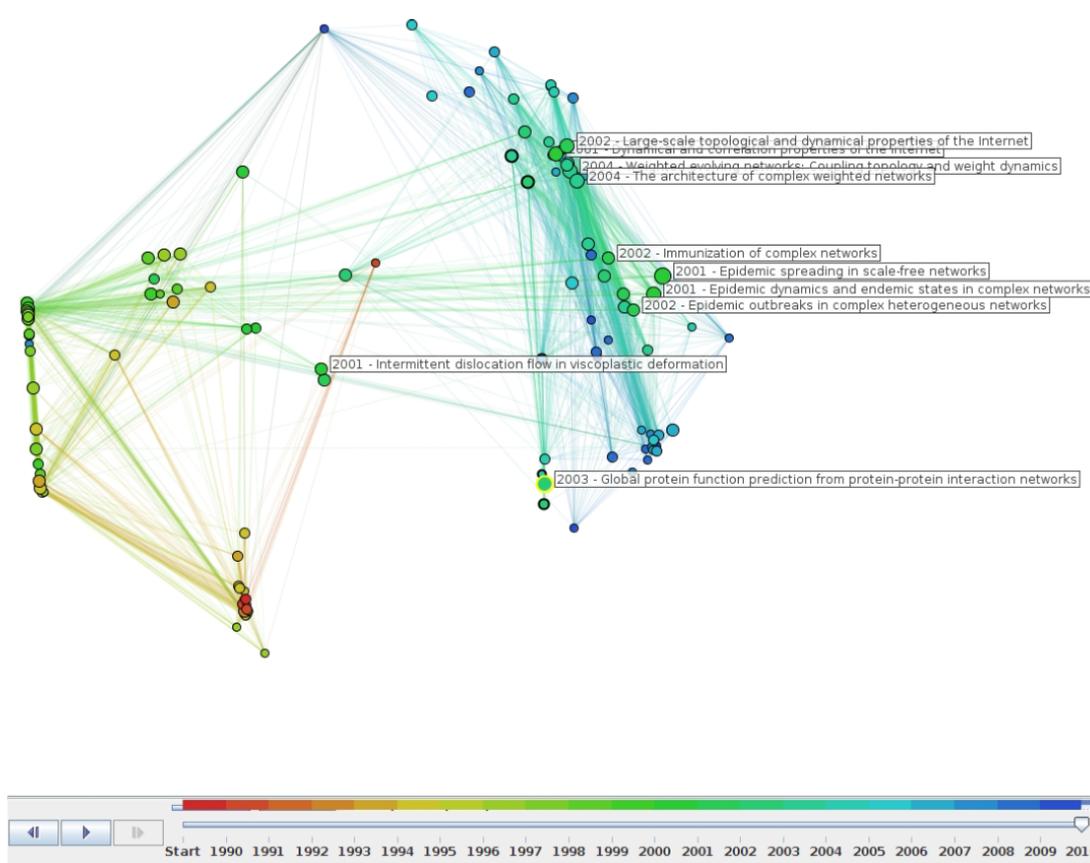


Figure 2.23 An example of node-link diagram maps attributes on nodes and edges using colour and size visual channels [2].

On the other hand, small visualisation glyphs, multivariate glyphs or radial plots are employed as substitutes for node representations to visually depict data associated with nodes. [200, 197, 104], while motif glyphs are employed to facilitate the understanding of network structure [50]. For example, Dunne and Shneiderman [50] employ cliques and fans glyphs as a means to replace common links and subgraphs, representing topological patterns within the network, but these are beyond the scope of our research. The research by Becker et al. [21] represents an early instance of employing the glyph approach. In their study, the thickness and colour were utilised on edges to visually represent the volume of incoming and outgoing telephone communication to provide a high-level view of network traffic across various significant cities in the United States, see Figure 2.24.

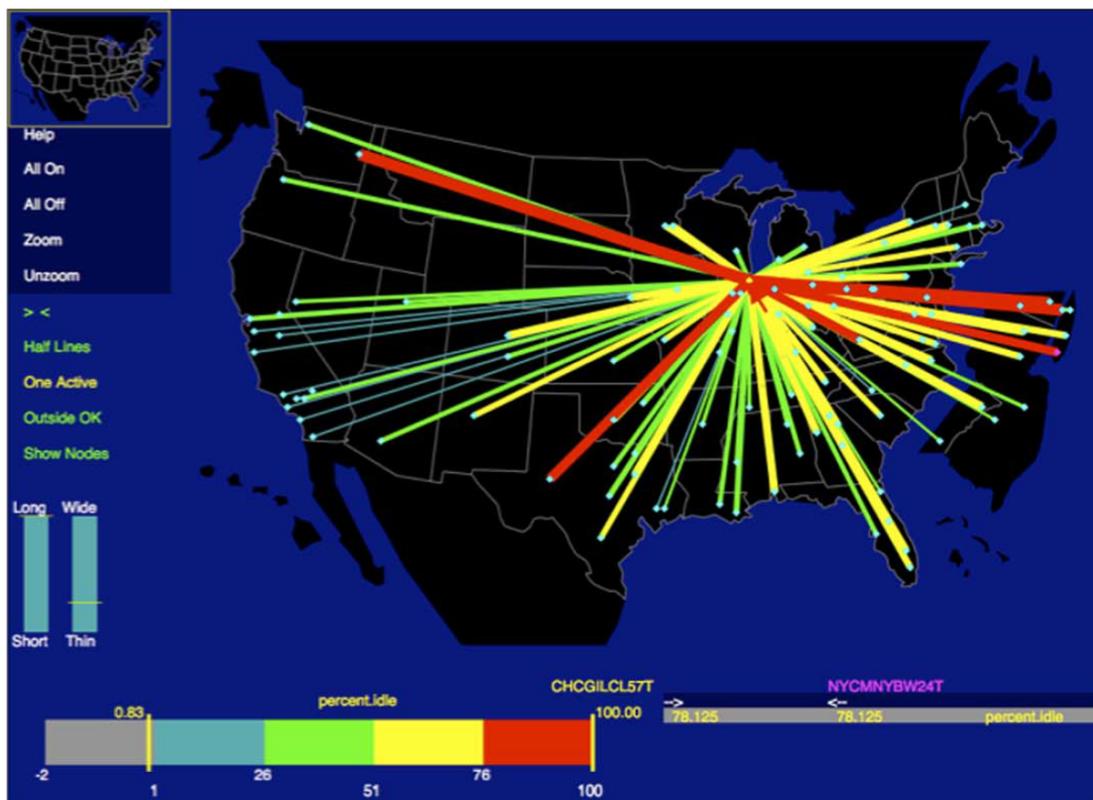
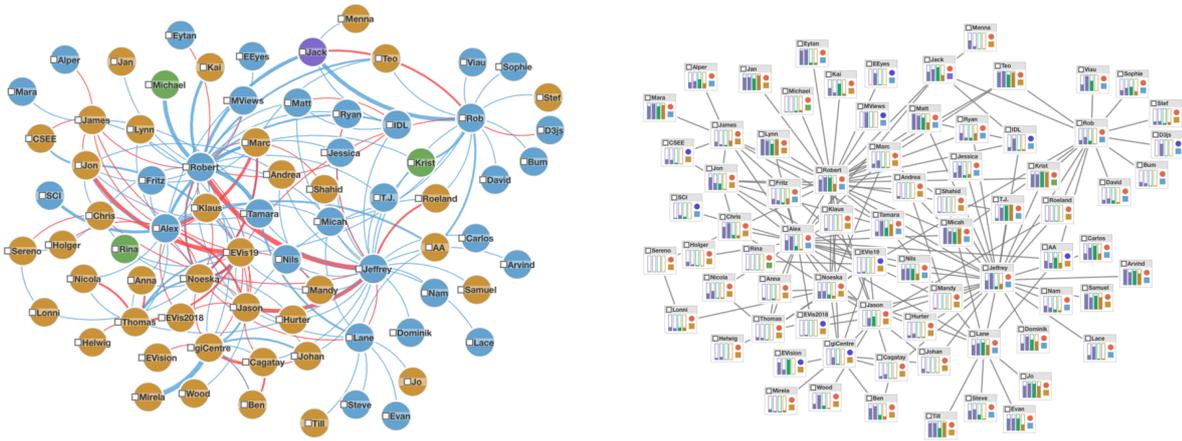


Figure 2.24 A node-link representation where nodes were placed according to their natural geographic location. Thickness and colour of edges encode to visualise network traffic [98].

In the context of matrix representations, it is possible to utilise colour coding or substitute cells with small icons to visually convey edge properties. Similarly, node attributes can be depicted by employing coloured node labels [104]. Nobre et al. [136] examined the use of on-node encoding in node-link layout and assessed the advantages and limitations of adjacency matrices and node-link diagrams in relation to various tasks performed on multivariate networks. Two variations of node-attribute visual encodings were developed based on the number of attributes, see Figure 2.25. For 1-2 attributes, the size of a circle and colour mappings are utilised for representing numerical and categorical values, respectively. To incorporate more attributes, nested bar charts and coloured glyphs were employed. Edge attributes were depicted through colour and thickness for categorical (specifically, edge type, which consisted of just two categories) and quantitative data (specifically, edge weights ranging from 1 to 5), respectively.



(a) Encoding two attributes on nodes and edges through two distinct visual mappings.

(b) Visually presenting multivariate data on nested bar chats as nodes and coloured glyphs.

Figure 2.25 a) Node size and colour were used to visually represent a categorical and a quantitative attribute, respectively. b) Multivariate data on nodes is represented by nested bar charts for numerical values, and coloured glyphs were used for categorical values [136].

Their decision to select visual mapping on nodes is based on the existing literature [133]. However, the authors did not present any design rationale for the selection of edge encoding.

The utilisation of bar charts overlaid on the edges [161] has been employed to encode multiple attributes in the context of node-link diagrams. Another approach is to employ multicoloured segments with varying line widths to depict different numerical properties of links [107].

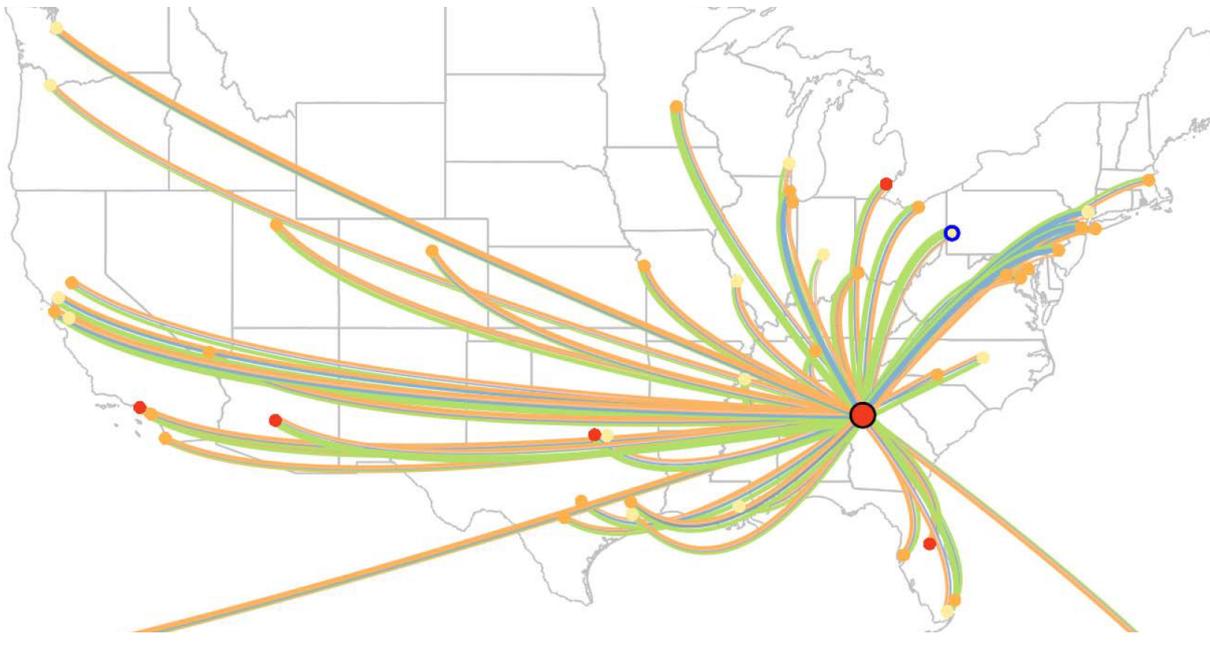


Figure 2.26 Several numerical attributes are encoded by the width of the coloured segments [107].

Ko et al. [107] proposed a new visualisation method called *multiple threads* for representing multiple numerical attributes of links. The approach employed in this study utilises multiple parallel links that consist of coloured segments with different thicknesses [107]. As shown in Figure 2.26, each link in the graph comprises multiple parallel threads, with the width

of each thread being adjusted proportionally to the value of a corresponding link variable. Nonetheless, challenges arise due to potential difficulties in colour discrimination, and this approach's scalability is limited.

Schöffel et al. [161] conducted a study on a methodology for representing multivariate data on graph edges (see Figure 2.27). Different sorts of bar charts were exhibited at the intersections of nodes. A study was done with a sample size of 89 participants to assess the impact of different types of bar charts on response time and accuracy in graph interpretation tasks. The study results indicated no statistically significant differences were observed in either response time or accuracy across the various types of bar charts evaluated. The subjective preference ratings revealed a preference for bars positioned on the edges rather than being centred around the edges. Additionally, bars of equal size were favoured over bars sized based on edge length. Furthermore, bars oriented orthogonally to the edges were favoured over their parallel counterparts. Nonetheless, it should be noted that the study did not provide statistical analyses to validate the discerned trends.

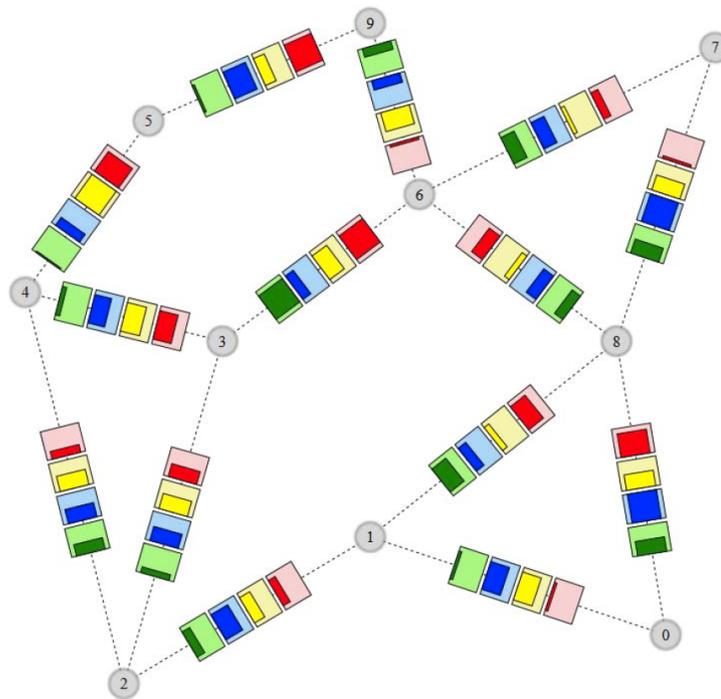


Figure 2.27 Link using coloured bars for encoding four quantitative edge attributes. This approach involves assigning specific colours to each attribute and representing their values through the length or height of the corresponding coloured bar. [161].

The existing corpus of literature regarding network visualisation methodologies is substantial and continuously growing. Numerous surveys and taxonomies have been conducted on diverse variations of multivariate graphs and their visualisation methodologies [104]. For instance, temporal graph visualisation design space [104] and multivariate graph visualisation [135] have been subject to thorough investigation. Numerous methodologies have been suggested to investigate the multiple attributes within a temporal multivariate network.

In a dynamic context, both the network topology and the values of attributes have the potential to change over time. Archambault et al. [6] provide a formalisation of multivariate temporal

networks, focusing on integrating time with other dimensions in the visual representations. The authors highlight three temporal aspects in their study, namely structure, behaviour, and evolution. Structure refers to the topology of the underlying network at a specific point in time, while behaviour pertains to changes in attributes. Lastly, evolution encompasses changes in the network's structure over time. Additionally, their model incorporates the concept of influence, which refers to a functional relationship between network elements and characteristics. For instance, an attribute denotes a graph-theory property that is obtained from the structure of the network.

In their study, Saraiya et al. [159] investigated the visualisation of multivariate networks by focusing on the representation of graphs connected with time series data. The authors explored different methods for indicating changes in the graph data at specific points along the timeline. The experiment involved the examination of a singular attribute, which was visually represented by utilising node colour. This representation employed a colour spectrum that spanned from green to yellow to red. Furthermore, the researchers conducted an investigation into multiple attributes by utilising nested views that incorporated heat maps and line charts as time series glyphs. These visualisations were employed to depict temporal patterns of nodes that are embedded within a static network structure.

Besides encoding attributes on data entities, in the literature, *attribute-driven layouts* consider including network attributes to calculate the layout, even if these attributes may not be explicitly visually represented explicitly [135]. For example, determining node placements in the graph layout can be achieved by calculating attribute similarity. Alternatively, nodes exhibiting similar attributes may be clustered together to diminish the overall size of the network.

As an example of the attribute-driven layouts, *PivotGraph*, as proposed by Wattenberg [204], presents a consolidated perspective of a multivariate graph. The approach involves combining nodes and edges that possess identical values for the designated categorical attribute value, such as node type, and afterwards utilising colour and size as visual representations of this information within a 2D grid-like node-link layout. The sizes of the points in the visualisation are directly proportional to the number of individuals who possess the corresponding value. The links between the points are represented by varying widths, which are also proportional to the number of edges linking the points.

An alternative illustration was found in the GraphDice system [24] in which multiple non-spatial attributes were encoded in two dimensions. Multivariate data was used directly define a layout by using a scatterplot for the nodes and superimposing edges onto this, and nodes are positioned according to nonspatial attributes. However, such specialisations are not in the scope of this study.

The prevalence of attributed graphs is on the rise, and a significant issue in attribute encoding is the effective representation of characteristics linked to nodes and/or edges inside the network [104]. Therefore, the illustration of node-link diagrams necessitates incorporating appropriate design elements for fundamental drawing primitives for both edges and nodes [197].

The analysis of the characteristics of node-link diagrams yields valuable insights that can enhance their comprehensibility in practical applications. However, incorporating additional information, such as detailed attributes of nodes and edges, poses a significant challenge in the

context of multivariate graphs. Employing multiple visual variables or multivariate symbols (or glyphs) is required to encode multiple attributes, which introduce an additional layer of visual complexity, imposing a considerable cognitive load when attempting to identify patterns and comprehend the visualised information.

Nobre et al. [135] focused on examining multivariate network visualisations. Two distinct encodings are distinguished in node-link diagrams: on-node encoding and on-edge encoding. While on-node encoding was found to be a favourable performance in graphs comprising diverse sorts of node attributes. They indicated that limited research had investigated the visualisation of edge attributes. They stated a lack of appropriate visualisation methods for this encoding type of bivariate or more attributes, presenting a novel avenue for future research.

Visualising edges in node-link diagrams has greater challenges than representing nodes, mostly due to the intrinsic constraints imposed by the narrow profile of the edge primitive. This limitation results in a reduced drawing space for edges, in contrast to the relatively larger space available for nodes. Therefore, it is an open question to employ efficient visualisation methods to facilitate the examination and analysis of both the network structure and the associated bivariate properties on its edges.

2.4 Uncertainty Visualisation

Existing techniques for visualising uncertainty frequently fail to effectively facilitate decision-making processes because they are restricted in their ability to depict uncertainty data accurately. Numerous studies examined in this section have underscored this constraint, thereby emphasising its significance and the practical need for enhanced visualisation methodologies for more reliable and comprehensible visual representation in the decision-making process.

2.4.1 Uncertainty

In its most basic definition, uncertainty is a result of a *lack of information* [89]. Uncertainty is frequently quantitatively represented by scalar values like probability, error, percentage, distance (e.g. from the true value), variance or standard deviation [69, 123] alongside the primary data or as an intrinsic part of the data, deduced from the data's description. Researchers have distinguished two types of uncertainty: *aleatoric* and *epistemic* [178]. While aleatoric uncertainty results from randomness, such as findings by chance, which are entirely objective in which outcomes vary each time an experiment is repeated. Epistemic uncertainty refers to a lack of knowledge, or in principle, the potential to lessen with further information, but this is generally not possible in practice.

Uncertainty can be present in various data sources utilised for information visualisation and visual analytics [23], and diverse forms of uncertainty are present in all facets of life. Weather uncertainties arise from an absence of accurate forecasts [95], network connection uncertainties arise from a lack of bandwidth reliability, and sensor data uncertainties arise from a lack of accurate or complete data [97]. Data uncertainty can have a significant impact on the analysis of

the data and the following decision-making process. Consequently, uncertainty should be taken into account when analysing and communicating visual data.

The concept of uncertainty is thoroughly established in several areas that engage with measured data. As an example, error bars are commonly employed to represent the uncertainty associated with measurements, demonstrating the standard mean of error or related descriptions of variability or uncertainty [138]. Nevertheless, comprehending the most standard forms of uncertain communication poses a significant difficulty for novices and experts alike [22]. This challenge arises, at least in part, from the abstract nature of uncertainty and ineffective communication approaches. It is crucial to acknowledge that the visualisation of uncertainty is a dual challenge, encompassing both technical and cognitive aspects. It requires a deeper understanding of the data and relationships within the graph and the ability to communicate and interpret the uncertainty to the user effectively. One prominent concern is developing visual metaphors that effectively build an intuitive cognitive connection between uncertainties and the data. This connection aims to enhance the user's understanding of the relationship and reduce the likelihood of overlooking important information. [69].

Uncertainty poses a significant challenge that cannot be disregarded, as failure to address it adequately may result in incorrect or imprecise decision-making [97]. It is a demanding and complicated notion, and its representation through visual means has become indispensable within the field of data science and analytics [122]. Given that the error is inherent to the data analysis process, a simple depiction of the data will not fulfil the intended objective. [97]. Therefore, it is imperative to employ effective visualisation techniques to satisfy the demands of the growing representation of primary data and its associated uncertainty, such as confidence and variance.

2.4.2 Uncertainty Visualisation

The challenge of uncertainty visualisation has been recognised as a prominent area of research in the field of visualisation. Over the years, hundreds of studies have been conducted in the visualisation community on this emerging topic [91]. Numerous survey papers have been published, offering comprehensive coverage of uncertainty visualisation from various angles. Pang et al. [141] present an influential work that applies a general taxonomy of visualisation techniques to the domain of uncertainty visualisation. Their framework classifies based on: input data values and associated value uncertainty; the data's position and its positional uncertainty; extent specifications for location and value; discrete versus continuous visual encoding; and axis mappings. This taxonomic approach brings valuable coherence for cataloguing the extensive array of general and uncertainty-focused visualisation approaches. The categories also assist technique selection based on data characteristics. Nevertheless, these taxonomies have limitations in elucidating how uncertainty visualisations function and guiding novel technique conception. Hence, the authors also classify uncertainty visualisation methods based on the subsequent categories: inclusion of glyphs, geometry addition, geometry modification, attributes modification, employment of animation, utilisation of sonification, and implementation of psycho-visual approaches.

A recent survey by Kamal et al. [97] categorised uncertainty visualisations, which is comparable to the classification approach by Pang et al. [141]. These categories include geometry,

attributes, animation (blinking, motion), visual variables, traditional graphical approaches (such as box plots, scatter plots, and histograms), and glyphs. Additionally, the authors provide a concise overview of the theoretical underpinnings of uncertainty visualisation and sources and models of uncertainty. They also discuss several methodologies for evaluating uncertainty visualisation techniques and provide future research directions in this field.

Graphical representations play a crucial role in effectively communicating facts and ideas across several domains. When examining uncertainty within the information visualisation framework, it becomes evident that it manifests itself at each level of the visualisation process. In accordance with the work of Brodlie et al. [29], a distinction can be made between *visualisation of uncertainty* and *uncertainty of visualisation*. When addressing uncertainty visualisation, the primary emphasis is typically placed on demonstrating the presence of uncertainty inherent in the data. The latter term refers to the further uncertainty that arises from visualisation, in addition to the uncertainty already present in the data. To be more specific, the uncertainty of visualisation refers to the uncertainty that arises during the stages after the data acquisition in the visualisation pipeline [29] (as shown in Figure 2.28). It is obvious that the uncertainty of visualisation is undesirable and is outside the scope of this thesis. Our research is centred on the initial notion, visualisation of uncertainty.

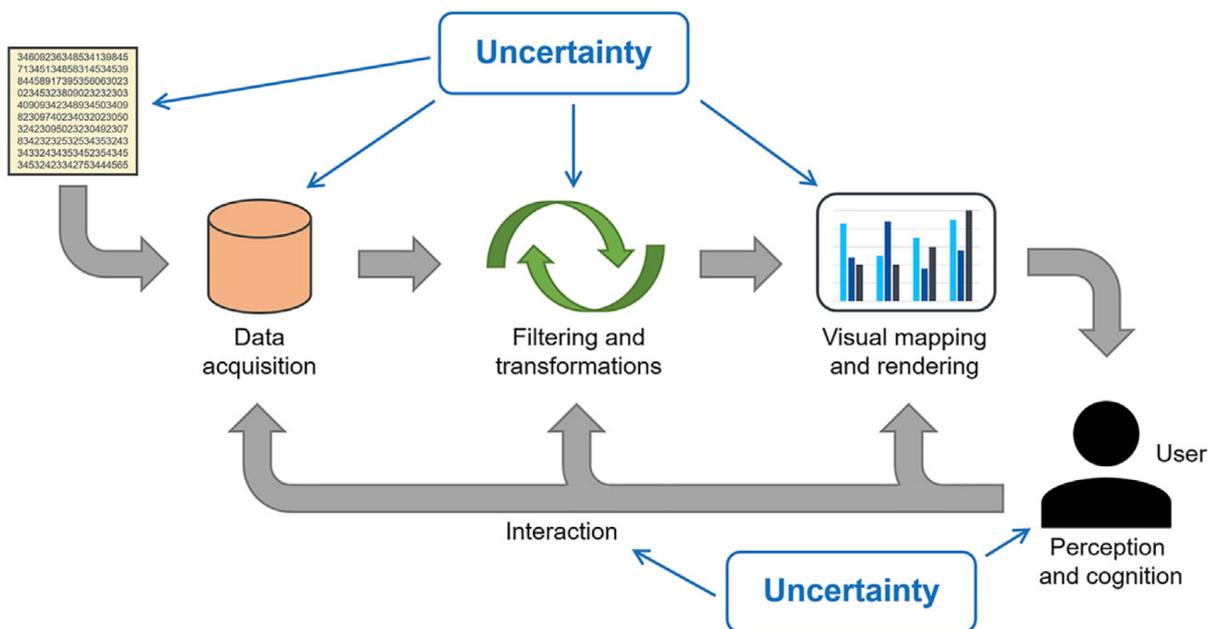


Figure 2.28 The visualisation pipeline can be extended to incorporate considering uncertainty at each stage. (Source: [205], p. 3)

Hullmann [88] focuses on the examination of existing uncertainty visualisations and highlights that varying designs have a significant impact on the outcomes. It has been observed that uncertainty visualisation can be susceptible to errors, particularly when incorrect inferences are drawn due to inadequate communication on the underlying statistical model. For instance, there may be confusion between standard deviation and variance.

A wide range of applications involve the utilisation of uncertain data, which may arise from several sources, such as inaccuracies, incompleteness, and inference. However, most graph

visualisation methods that have been presented focus just on visualising the graph itself, without incorporating the related uncertainty data [197]. In their study, Brodlie et al. [29] examined uncertainty-related challenges. They identified several key issues, including the complexity of uncertainty, the diverse ways in which uncertainty information is presented, the propagation of uncertainty, the introduction of an additional dimension to visualisations, the tendency for uncertainty to overshadow certainty, and the interdisciplinary nature of uncertainty.

Several visual variables have been proposed for representing uncertainty, including "Bertin's retinal variables" [23], blur, and transparency [121]. MacEachren et al. [122] examined the intuitiveness and performance of different visual variables, such as fuzziness, location, and saturation, in representing uncertainty for single objects during map reading tasks. In contrast to our study, they evaluated using point symbols rather than lines. The participants were asked to evaluate the level of intuitiveness of a predetermined group of visual variables, see Figure 2.29. Nevertheless, it should be noted that not all visual variables are effective for illustrating uncertainty regarding their intuitiveness. Participants favoured fuzziness and location, whereas saturation received a very low ranking. [122]. They also evaluated iconic symbolisation for representing uncertainty. The theory of visual semiotics of uncertainty [122] has significantly influenced the development of many uncertainty visualisation applications. Guo et al. [71] refine these visual variables regarding graph edges.

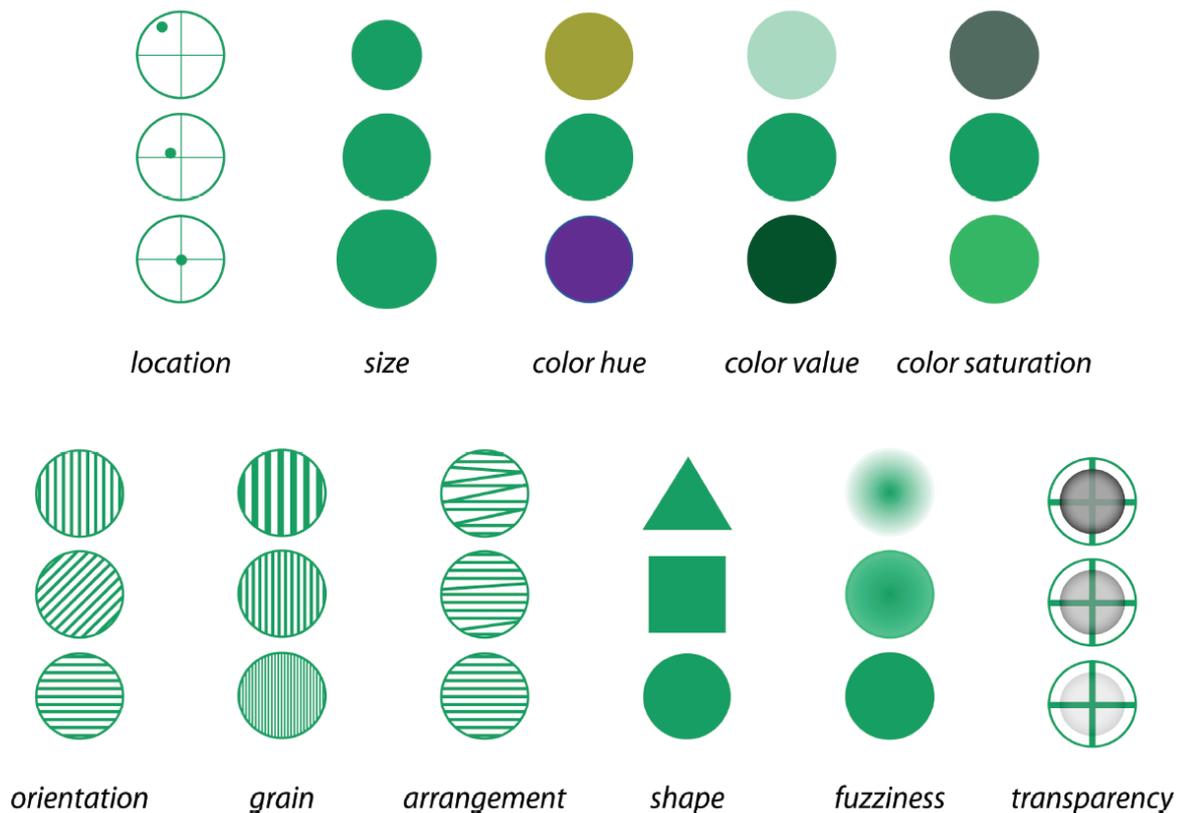


Figure 2.29 Eleven visual channels have been examined for representing uncertainty on point symbols [122].

2.4.3 Uncertainty Visualisation in Graphs

Graph data displays three distinct forms of uncertainty [100, 197]. Initially, a node's existence may be subject to uncertainty. Additionally, uncertainty may be present in the links between nodes. Furthermore, the attributes associated with both nodes and links can also be subject to uncertainty. However, it is important to note that the placement of visualised nodes is not an inherent uncertainty, except for geographic map layout nodes, which have location uncertainty in the real world since the graph description or a graph-drawing technique determines it. Our research focuses on representing uncertainty attributes on both nodes and edges. In this context, the term uncertainty refers to quantified uncertainty that can be represented visually.

The incorporation of graph visualisation with uncertainty has attracted considerable attention in recent years. Uncertainty is commonly represented as a visual variable [71, 122]. This visual variable is then applied to a specified graph layout.

Within node-link diagrams, lines serve as ubiquitous graphical elements. Information can be encoded using these lines through the modulation of a distinct attribute, such as grayscale, in accordance with its quantitative or qualitative value. Incorporating uncertainty information into visualisations through lines introduces an augmented avenue of communication. As an illustration, uncertainty could be represented by lighter lines, whereas darker lines could denote certainty. Opacity, wavelength, blur, colour, width, and sketchiness are examples of attributes assigned to lines that are found in the literature.

Boukhelifa et al. [27] conducted a study to assess the level of intuitiveness associated with the line attributes of *sketchiness* as a new visual variable in the context of visualising uncertainty. Grayscale, blurred, and dashed lines have been compared to sketchy lines, see Figure 2.30. The term "sketchiness" pertains to deviations observed in a line that imitate the characteristics of hand-drawn lines. These deviations exhibit a higher level of irregularity compared to wave-like patterns. As the magnitude of deviations increases, so does the level of sketchiness. While the design of sketchiness may be considered unprofessional, it has been found that accurately perceiving the degree of blur and grayscale is also challenging.

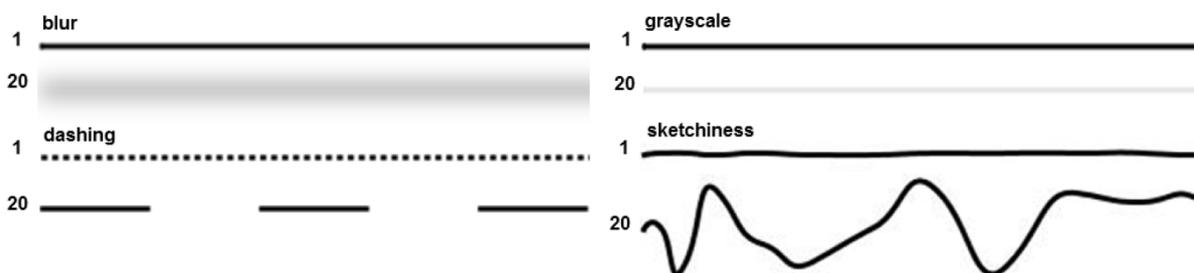


Figure 2.30 To determine which visual designs individuals preferred concerning uncertainty encoding, they requested that participants choose one of four possible options [27].

The study conducted by Guo et al. [71] focused on the utilisation of various combinations of visual variables in order to represent the uncertainty associated with graph edges, see Figure 2.31. The researchers assessed the user's perception of undirected edges that represented two characteristics simultaneously, namely, strength and certainty. Various combinations of visual variables

were evaluated for distinct tasks. The researchers chose to utilise width, hue, and saturation as visual representations of the strength (primary value) edge attribute. Additionally, to encode the uncertainty attribute, they selected lightness, grain, fuzziness, and transparency. It was found that the perceptual distinction between lightness and saturation is significantly greater than that between lightness and hue. The findings of their study suggest that the combination of brightness and hue does not yield favourable outcomes, while the integration of fuzziness and width is not advisable.

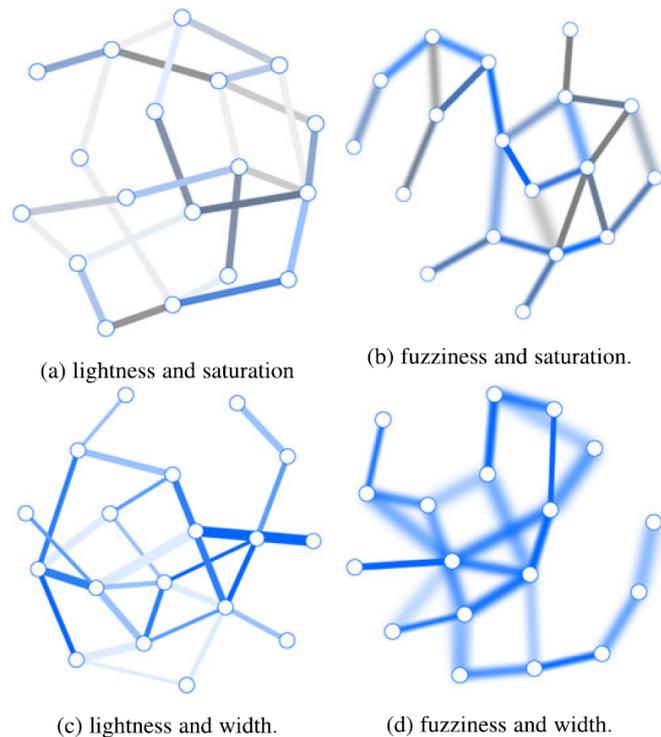


Figure 2.31 Guo et al. [71] conducted a user study of paired visual variables to visualise uncertainty as an additional dimension in conjunction with the primary attribute of the edge.

In their empirical study, Bae et al. [13] examine several visual encodings pertaining to causality, strength, and uncertainty. The objective of their research is to identify the most effective representations and offer recommendations for the design of causal diagrams, see Figure 2.32. The focus lies in the empirical evaluation of several attributes pertaining to the visualisation of causality. The visual cues that were assessed encompassed the use of arrows versus tapered lines to indicate the direction of causality, the utilisation of width versus hue to represent the strength of causality, and the consideration of brightness versus fuzziness versus granularity to convey the level of uncertainty. The researchers reached the conclusion that the indication of causality direction can be achieved by employing arrows or tapered lines. The visual attributes of width, brightness, and fuzziness are suggested for representing strength and uncertainty. They stated that visual representations employing width and brightness demonstrate comparable performance to representations utilising tapered-number-number depictions.

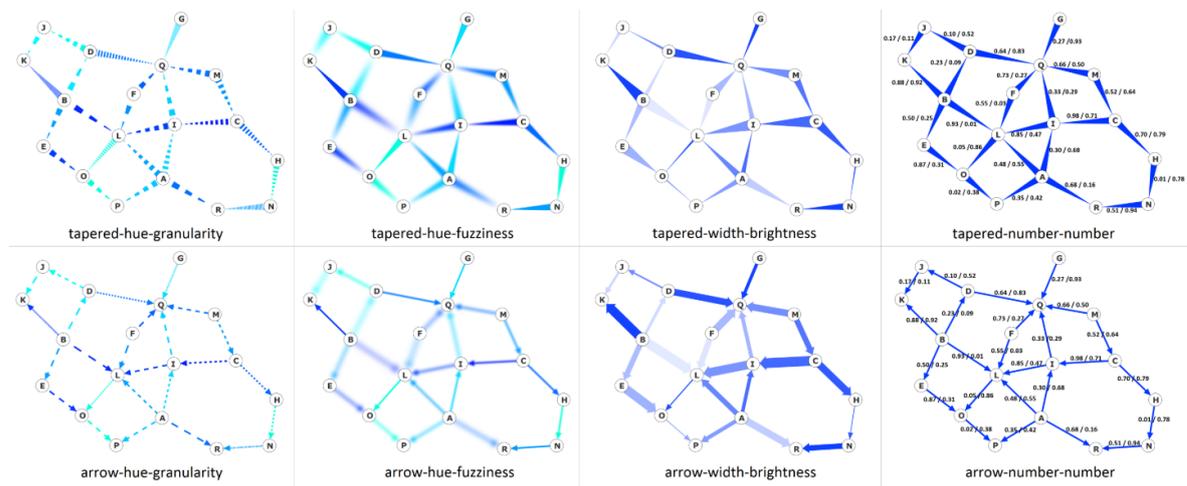


Figure 2.32 Eight different combinations of visual variables used in their experiment [13].

2.4.4 Summary

Uncertainty visualisation approaches can help highlight the uncertainty, facilitating a comprehensive portrayal of the investigated data. [138]. It is essential to recognise that a universally applicable approach to uncertainty visualisation, capable of enhancing decision-making across all domains, does not exist. Furthermore, there are no assurances that the presentation of uncertainty to readers will invariably lead to improved judgements or increased trust [139]. Hence, it is imperative for visualisation designers to exercise meticulous consideration for every design option they make, as failure to do so may increase the complexity of the decision-making process.

When it comes to depicting uncertainty on edges, the visual encoding channels are employed to determine the visual attributes of line marks, including colour, size, and transparency [133]. Encoding channels offer the advantage of modifying an existing mark, such as increasing its transparency when there is significant uncertainty. Hybrid techniques can be devised by combining marks and encodings to convey uncertainty. Nevertheless, a constraint exists regarding the number of visual channels that can be utilised to depict bivariate data on edges. It is worth mentioning that many colour channels, including hue and saturation, possess perceptual integration, hence complicating their use in bivariate data analysis [43].

In this subsection, various methods for uncertainty visualisation have been explored in the context of node-link diagrams. Throughout discussions on uncertainty visualisation, it is clear that many state-of-the-art methods focus on presenting uncertainty in isolation or without proper integration with primary data attributes. This thesis introduces the Vizent edge design as a novel solution, which provides a dual-encoded approach to visual representation by embedding uncertainty directly into the visual representation of network edges. Unlike traditional methods that often separate uncertainty from primary data values, the Vizent edge design seamlessly combines these elements, thereby enhancing the graph's interpretative value. This approach addresses a critical gap identified in the literature, where visualisation techniques fail to adequately convey the relationship between primary data attributes and their underlying uncertainty. These advancements are designed to improve decision-making processes, allowing users to make more informed decisions based on a comprehensive view of the data presented.

Comparative analysis with traditional methods has demonstrated that the Vizent design not only increases the accuracy of data interpretation but also enhances user understanding, as validated by the usability studies detailed in Chapter 5.

2.5 Graph Summarisation

Network data is everywhere, encompassing various forms such as e-mail traffic between individuals, telecommunication networks, transportation networks, and financial networks [197]. Although graph visualisations predominantly try to offer insight into diverse patterns underlying the connections among complex and large data elements and their distinct attributes, it is unfortunate that they suffer from a lack of scalability. As the number of nodes and edges grows, the complexity of the graph increases while legibility diminishes. [126, 112].

Node-link diagrams frequently suffer from *visual clutter* [117, 181, 52] when drawing datasets of a large scale [67], which can negatively impact usability, aesthetics, and the interpretation of data. Furthermore, overlapping between nodes, edges, and node-edge contributes significantly to the visual clutter [144]. According to Rosenholtz et al. [156], the presence of *visual clutter* in data visualisation can be identified in the following manner:

“Clutter is the state in which excess items, or their representation or organisation, lead to a degradation of performance at some task.”

The task of representing large graphs through node-link diagrams poses significant challenges within the domain of information visualisation and graph drawing communities. The techniques that are considered most effective for automatically drawing graphs are primarily designed to provide aesthetically readable layouts tailored for moderately sized and sparsely interconnected graphs. Despite the considerable research conducted in graph visualisation, conventional node-link diagrams only effectively display graphs with a restricted number of nodes, often ranging up to a few hundred.

Users could closely examine and analyse specific local aspects within the displayed content through the use of interactions like zooming and panning. Nevertheless, it is important to note that this particular method of interaction exhibits several inherent constraints, including inadequate navigation patterns and potential user disorientation [84].

Focus+context, a well-known interaction strategy, can be implemented as an alternative category of interaction methods to allow observers to simultaneously obtain an overview of the entire network and view parts of the main interest shown in full detail. However, well-known examples of focus+context techniques for investigating large graphs, such as traditional *fish-eye views* [62], impose significant distortions that frequently result in diminished legibility of paths and other noteworthy structures [199].

Numerous graph reduction methods have been suggested to reduce the graph size so that important structures can be perceived easily, including clustering [16], edge filtering [93], and sampling [208, 214]. Edge-bundling algorithms [64, 184] (see Figure 2.33), and advanced layouts [8, 215] were developed to enhance the spatial arrangement of nodes and edges in order to minimise visual clutter. Despite the graph’s manageable size in terms of screen space,

effectively conveying the required information to users remains a challenge due to the limitations of human perceptual abilities, which often favour smaller graph sizes.

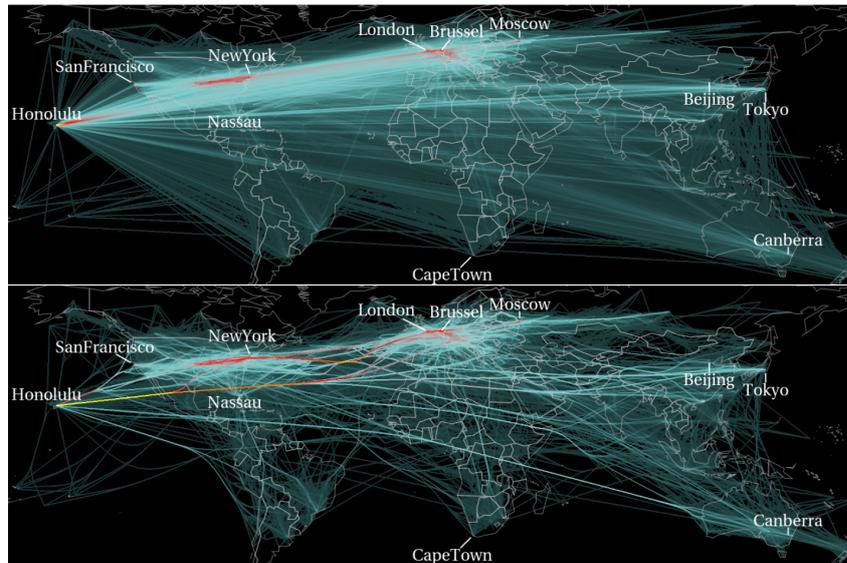


Figure 2.33 The layout of the Internet graph (top) has a total of 149661 edges. After implementing the edge bundling method, the resulting image is at the bottom; however, the result is still a hairball cluttered with edges. [64].

Graph bundling is a prominent visualisation technique that primarily emphasises graph edges in order to manage large and dense graphs. One approach to mitigate the issue of edge clutter in network visualisations is using edge bundling techniques. Edge bundling is a technique that groups links inside a graph based on predetermined criteria. These techniques, as illustrated in Figure 2.33, aim to aggregate groups of edges or portions of edges (i.e. their pixels), either implicitly or explicitly [198]. Through this process, they enable the visualisation of higher-level flow patterns within the network that would otherwise remain obscured. In the initial iteration of these algorithms [184, 83, 64], without accounting for their integration with other methodologies, the primary advantage was limited to the reduction of visible items. They tried to reduce the space dedicated to the links, increasing the possibility of making nodes visible. As opposed, the aggregation of links in a traditional way led to a more ambiguous visualisation [12].

The scale of graphs has been expanding as a result of the ongoing accumulation of data. As a consequence, it becomes progressively more challenging to extract valuable and easily understandable insights from large-scale graphs. Although data summarisation approaches have been researched comprehensively, only recently has summarising interconnected data, or *graphs*, gained attention [119]. Graph summarisation is a technique that is gaining prominence as a potential solution to address this particular issue.

Graph summarisation can generally serve to eliminate the noise and identify underlying patterns within the data, even though the concept of graph summarisation lacks a precise definition [119]. Graph summarisation is a technique used to create smaller graphs that use significantly less storage space while retaining the distinct characteristics of the original, large-scale graphs [119]. This particular representation exhibits a reduced scale, hence facilitating the comprehension process for observers. Graph summarisation studies can be classified into

summarisation of static [105, 151], dynamic [165, 183], weighted [118, 73], or attributed graphs [212, 106]. However, most summarisation algorithms function on static networks, utilising the structure of the graph and, if available, the properties of the nodes and edges.

The process of graph summarisation enhances the efficiency of graph analysis by generating a compact representation of the graph that may result in some loss of information but facilitates a more streamlined analysis [151]. The literature has extensively examined a diverse range of strategies, each characterised by a distinct approach. Among the techniques utilised by researchers are *grouping or aggregation-based methods*, *simplification-based methods*, *compression-based methods* and *influence-based methods* [119].

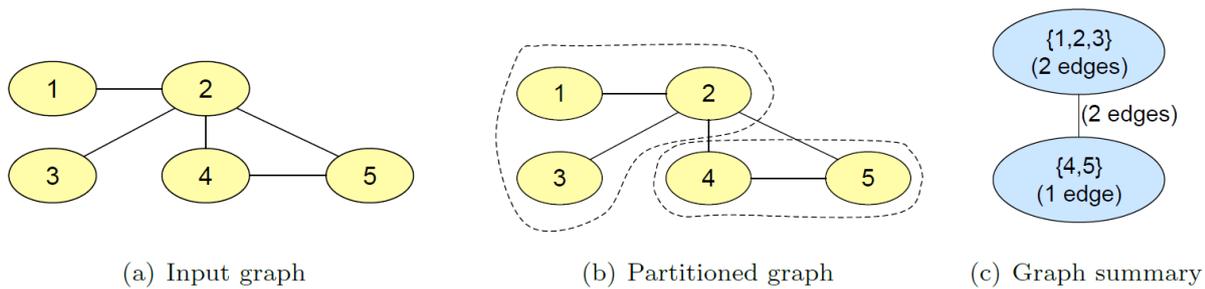


Figure 2.34 The example of graph-structure summarisation [114] is centred on graph topologies and does not incorporate attribute values.

One of the most utilised grouping-based methods, such as edge-grouping or node-grouping, is generating a supergraph. A supergraph is a graph in which nodes are iteratively combined into *supernodes* (a set of vertices), and edges are aggregated into *superedges* (each superedge holds information about how many original edges it represents) [114], see Figure 2.34.

Simplification-based graph summarisation involves eliminating nodes or edges deemed less "important" in the graph, leading to the creation of sparsified graph. When comparing supergraphs, it is observed that a summary is comprised of a subset of the original set of nodes and edges, for instance, see Figure 2.35. Similarly, a concept, known as *backbone identification*, serves the purpose of discerning the most important edges inside a network and eliminating all other edges [26]. Backbone identification is particularly valuable when dealing with networks that are very dense, for example, where individual nodes and edges can no longer be seen, and the layout resembles a big hairball [26]. There are different approaches to identifying backbones. The simplest approach is to use node and/or edge properties to delete links that are less relevant. For example, *DrL* [125] layout algorithm manages to layout large and dense networks by only keeping the top n highest weight edges per node.

Link reduction is a straightforward method that involves selectively displaying edges that have weights exceeding a specified threshold or that meet particular criteria. Thus, solely those boundaries that the user may find interesting are illustrated. A further extremely straightforward solution could be to remove every node or every link. Although the decision to display nodes or edges exclusively may be subjective and yield debatable outcomes, there exist more advanced techniques that offer well-defined justifications for minimising visual complexity.

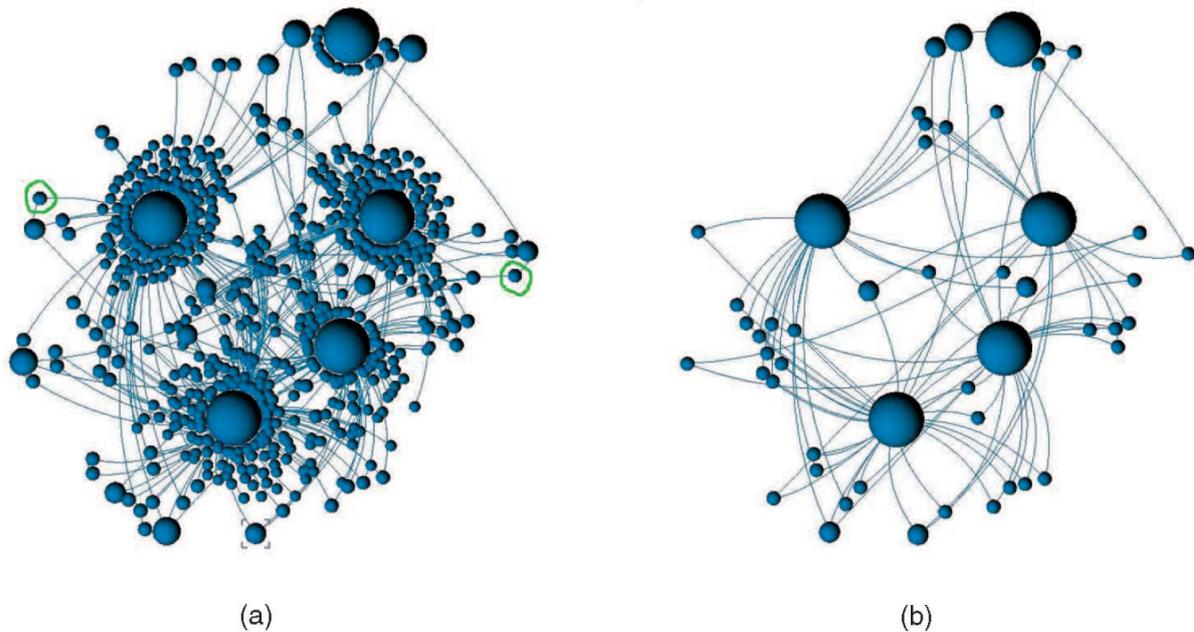


Figure 2.35 a) The original network is visualised using the node degree information. b) The graph is simplified through the removal of structural information, namely by eliminating one-degree nodes in order to reduce the visual complexity. [169].

Previous research [75, 187, 55] has also explored the issue of compressing large graphs, with a particular focus on compression techniques for WWW (Web) graphs. However, the primary objective of these graph compression methods is to achieve compact graph representations that are simple to store and manipulate by implementing encoding techniques like bitmap compression, and reference encoding, whereas graph summarisation methods [111, 186] concentrate on generating concise and comprehensible summaries.

Bit compression-based methods compress the original graph into a smaller, more manageable summary by employing techniques from data mining. In order to generate the summary or model, the number of bits necessary to represent the original graph is minimised. As a consequence, the summary produced is significantly more compact than the original graph; however, it still contains essential information that facilitates a more comprehensive comprehension of the original graph's structure, including network structural patterns such as bipartite subgraphs.

There is a small number of influence-based summarising approaches. By utilising structural and node attribute similarities, the influence or diffusion process in a large-scale network can be summarised. [170].

There is still some disagreement within the literature over the appropriate structure and format of a graph summary. It is always application-dependent and can accomplish a wide range of tasks, including preserving the responses to graph queries, discovering new graph structures, merging nodes into supernodes, and merging edges into superedges, among countless other purposes. Nonetheless, the challenges encountered by individuals involved in the creation of graph summaries are commonly experienced throughout various fields [119]:

- The complexity of splitting or merging edges and nodes in a graph increases as the number of properties associated with the graph increases. This can lead to complications as real-world domain networks often exhibit an increasing heterogeneity of nodes [119].
- Graph summarisation's primary objective is to minimise the complexity of the input graph in order to facilitate subsequent analysis. Nevertheless, methods employed to accomplish this face the obstacle of processing substantial amounts of data.
- The primary outcome of a graph summarisation procedure is to retrieve the most interesting observations presented in a graph. However, determining what is deemed as *interesting* is frequently a matter of subjectivity, contingent upon particular objectives of the analysis.
- Determining the threshold of "sufficiently many results" will remain a challenge even after a successful elucidation of the criteria for defining what is of interest during graph summarisation. One example of a compression technique is examining the extent to which the size of a graph has been decreased in terms of raw bits. As an additional illustration, a simplification technique will examine the disparity in the total number of nodes and edges prior to and subsequent to the process of summarisation.

Generally, each technique is characterised by a distinct set of criteria that govern the determination of successful outcomes. Furthermore, it is worth noting that the absence of a ground-truth answer typically challenges each technique. In conclusion, it is advisable to try various approaches, evaluate their outcomes, and ultimately choose the most suitable one since there is no one-size-fits-all method.

2.5.1 Summary

This subsection briefly presented the graph summarisation techniques in the context of information visualisations. We discussed some graph reduction methods, interaction techniques, and simplification-based graph summarisation methods related to the work presented in this thesis. There are several methods for summarising graphs, each with advantages and limitations. However, selecting the suitable summarisation method depends on the application's specific requirements and requires a good understanding of the characteristics of the data and the desired level of detail in the summary.

In summary, graph summarisation is crucial for condensing large networks into more compact representations while preserving both structural and attribute information. However, existing approaches often focus more on reducing the visual complexity of graphs without fully addressing the retention of "more important" information pertinent to the user's particular needs. To overcome this, Chapter 3 presents an entropy-based graph summarisation algorithm that effectively balances complexity reduction with the preservation of important information. This algorithm applies information-theoretic principles to selectively prioritise and display graph elements based on their informational value, ensuring that key insights remain visible even in significantly reduced formats. Unlike traditional methods that may rely on heuristic or arbitrary criteria, our proposed algorithm utilises Shannon entropy to ensure that the most informative

subset of edges is maintained in the summarised outputs while accounting for the inherent uncertainties arising from the variability or unpredictability of edge information in the graph. Additionally, the algorithm's capability to dynamically adjust to user-defined parameters allows for customised summarisations, making it highly adaptable for various specific analytical needs and scenarios. The implementation and efficacy of this summarisation approach are thoroughly explored in Chapter 3, demonstrating the practical benefits and efficiency of the algorithm in real-world scenarios.

Chapter 3

Graph Summarisation Method

Contents

3.1	Introduction	54
3.2	Motivation	55
3.3	Facebook Datacentre Topology	56
3.3.1	Facebook's Fabric Network Design	57
3.3.2	Traffic pattern inside Facebook's datacentre network	57
3.4	Concept of the Variance in edge weights	59
3.4.1	Edge Variance	59
3.4.2	Applying Edge Variance Concept to Facebook Traffic Patterns	60
3.5	Information Theory	61
3.5.1	Shannon Entropy	61
3.6	Entropy-Based Graph Summarisation Algorithm	64
3.6.1	Overview	64
3.6.2	Implementation	65
3.7	Testing with Facebook network data	69
3.8	Limitations	73
3.9	Future work	73
3.10	Summary	74

Overview

Representation of all data is not possible due to the limitations of the canvas size, the cluttering, or the impossibility of interpreting the data by the user. Moreover, when the number of presented nodes and edges increases, the user performance decreases, and end-users take more time to distinguish visualised nodes or edges since the cognitive demands of such visualisations surpass the capabilities of the average human brain.

This chapter introduces a method designed to simplify complex graphs by applying real Facebook traffic patterns; however, this approach can be applied to various network datasets. Consider a network to be a complex system of roads, with traffic flowing at varying speeds and volumes. Similar to how traffic analysts seek to identify roads with the most fluctuating traffic patterns to improve flow and reduce congestion, our method identifies the most variable connections in the network by providing a simplified yet insightful view without getting overwhelmed by graph complexity.

3.1 Introduction

Graphs are commonly employed as a means of representing and analysing real-world entities and their interconnections. Graphs in numerous applications sometimes exhibit a substantial scale, characterised by an extensive number of nodes and edges, reaching into the thousands or even millions [186]. While visualising small, static networks is relatively easy, encoding a large or complex data set into a single visual representation might result in a cluttered and overcrowded display that prevents the user from fully comprehending its structure and contents [196].

Real-world graphs of ever-increasing size contain additional information, known as labels or attributes, beyond just the connections between nodes. This additional information can provide valuable insights into the underlying structure of the graph. In practical applications, visualising networks, even those of moderate size, can provide challenges due to issues such as overlapping elements and information loss. Moreover, incorporating additional variables into these visualisations can further complicate the task [197]. Clearly, making sense of all this information visually to facilitate comprehension poses a distinct problem.

End users generally find themselves immersed in vast information while facing constraints regarding their cognitive capacity and temporal resources for assimilating new information. The issue commonly referred to as "information overload" is a prevalent challenge observed in diverse fields. The phenomenon can lead to a state of *information paralysis*. There is a growing abundance of information available, yet there is a noticeable decline in the quality of the information being disseminated. Consequently, users encounter difficulty in discerning the relevant information. For example, individuals engaging in visualisations typically interpret just a limited portion of the available information at any one moment. *Cognitive load*, as seen in the field of cognitive psychology, is derived from the limited capacity of human short-term memory, which is commonly known to be able to retain a relatively small amount of knowledge, often around *seven plus or minus two* pieces of information [130]. Hence, it is crucial to minimise the

relevant information shown at any given moment to ensure ease of comprehension regarding the primary attributes of data characteristics.

Finding a solution to the issue of information overload necessitates the acquisition of attention, the provision of ways for more effective summarisation and filtration, the exploration of potentially relevant information, and the enhancement of information quality. In order to effectively mitigate the occurrence of misunderstandings, misinterpretations, and the inappropriate utilisation of information, it is crucial to take appropriate measures. One effective strategy for mitigating visual overload is to limit the amount of information that can be presented simultaneously. While maintaining a perceptible and comprehensible simplified depiction of the large and complex data set, the observer does not experience a sense of being overpowered by the entirety of the data.

3.2 Motivation

Visual analysis of large and complex networks has garnered increasing attention, not only among many research communities (beyond those solely focused on graph theory) but also among the wider public. Visual search tasks are frequently encountered in graph visualisations, where the objective is to identify specific edges or nodes. One primary reason for this interest lies in the pursuit of objective insights regarding network connections, aiming to monitor and comprehend them and subsequently implement measures to enhance or optimise their use. Nevertheless, it is worth noting that node-link diagrams have limitations when it comes to accommodating a large number of nodes and links, often resulting in *hairball-like* visualisations. The large network visualisations might be daunting and perplexing for experts and non-expert users. The presence of hairball-like visualisations hinders the ability to explore or analyse networks effectively, resulting in a lack of insights or, in certain cases, the formation of erroneous conclusions due to visual clutter. Therefore, users require a concise visualisation that effectively communicates the relationship between the structure of a network and its associated attributes.

Much work has been done on developing visualisation techniques that try to scale to generate visual representations of large graphs [30]. Most approaches aim to provide users with a comprehensive perspective of the complete graph. This commonly entails the utilisation of efficient layouts, multiscale clustering methodologies or matrix-based network visualisations [192]. While the objective of offering a comprehensive outline of the graph is a laudable goal, there are numerous situations where the user's focus is not on obtaining a holistic perspective of the entire graph but rather on addressing a specific task or problem associated with the graph.

Exploration of the visualised large and complex graph-based data through existing visualisation schemes in a short period becomes quite challenging due to exceeding the limits of short-term memory. Moreover, our ability to comprehend those graphs remains constrained, and we fail to absorb all the information as decision-makers' time is limited. In these cases, even with dynamic deployment, understanding and communication of these graphs without implementing a suitable data reduction method would impose a significant cognitive load on the viewer to understand the most critical aspects of the graphs due to the limited capacity of the human brain.

These challenges can be addressed by leveraging the user’s pre-existing knowledge of the tasks and selecting an appropriate summarisation technique.

Graph summarisation can be understood as a process of reducing the complexity of a graph while preserving important information. Our motivation is to identify the most informative features and patterns in graph-based systems in order to address information overload. Information theory has been applied to determine the relevant data to be displayed. One way to achieve this is by using entropy-based methods, such as Shannon entropy, which can quantify the amount of information in a graph. As a result, information theory can be valuable in highlighting the intended insights while hiding less “interesting” features in this pursuit. These insights, in turn, can be leveraged to develop more effective and efficient visualisations for human perception.

3.3 Facebook Datacentre Topology

Facebook is one of the leading web services providers in the world, and its network is made up of different datacentre regions with a WAN backbone connecting these regions [129, 56]. Facebook caters to a monthly user base of 2.23 billion people and has twelve datacentres spread across different locations. These datacentres employ various generations of datacentre network design [129]. Each region consists of at least one datacentre building, which is referred to as a *datacentre*. Facebook utilises two different intra-datacentre network designs on its backbone network: cluster-based design (older) and the state-of-the-art fabric design [5], as discussed in Section 3.3.1.

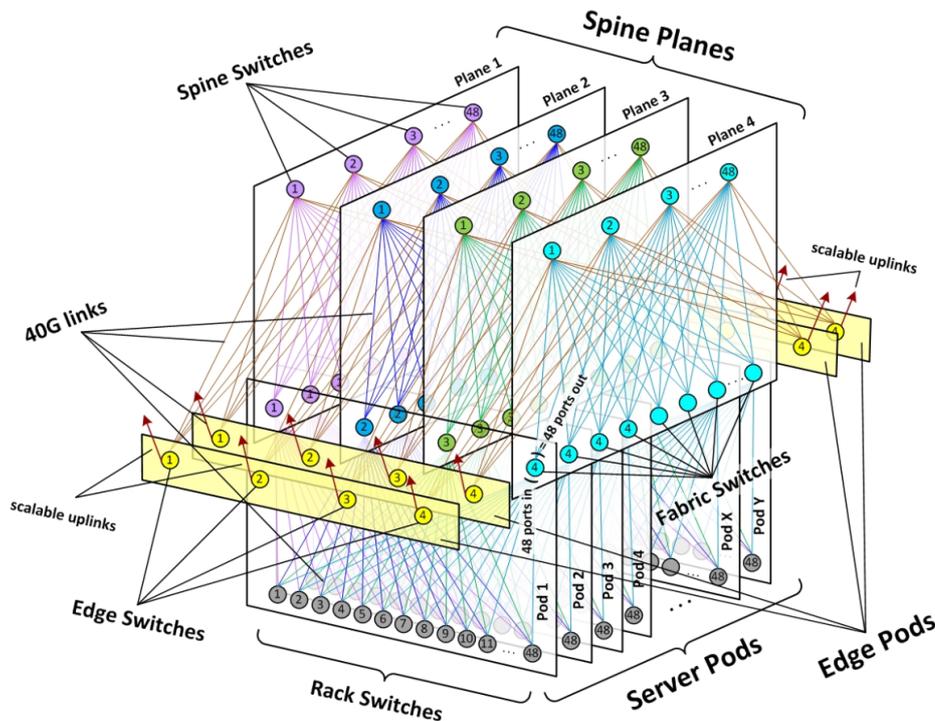


Figure 3.1 Fabric network design of Facebook data centre, which accommodates a large number of interconnected devices and a more complex topology [5].

3.3.1 Facebook's Fabric Network Design

This section provides an overview of the Facebook data centre and its connectivity, see Figure 3.1. An examination of the anatomy of the Facebook fabric design, as depicted in Figure 3.1, reveals four independent "planes" of spine switches, each of which can accommodate up to 48 independent devices per plane. The spine plane is responsible for managing inter-cluster traffic within the data centre. Each pod is served by a set of four devices known as *fabric switches*. Four fabric switches handle inter-rack traffic within the cluster, known as pod, and traffic incoming or leaving the cluster. Each pod consists of only 48 server racks (top-of-rack switches, TOR), and this form factor is always the same for all pods, see Figure 3.2. Each rack has only one top-of-rack switch with 10-Gbps Ethernet downlinks in which server machines are organised into the rack. Each fabric switch of each pod connects to each spine switch within its local plane. The combination of pods and planes creates a modular network topology that supports hundreds of thousands of 10G-connected servers.

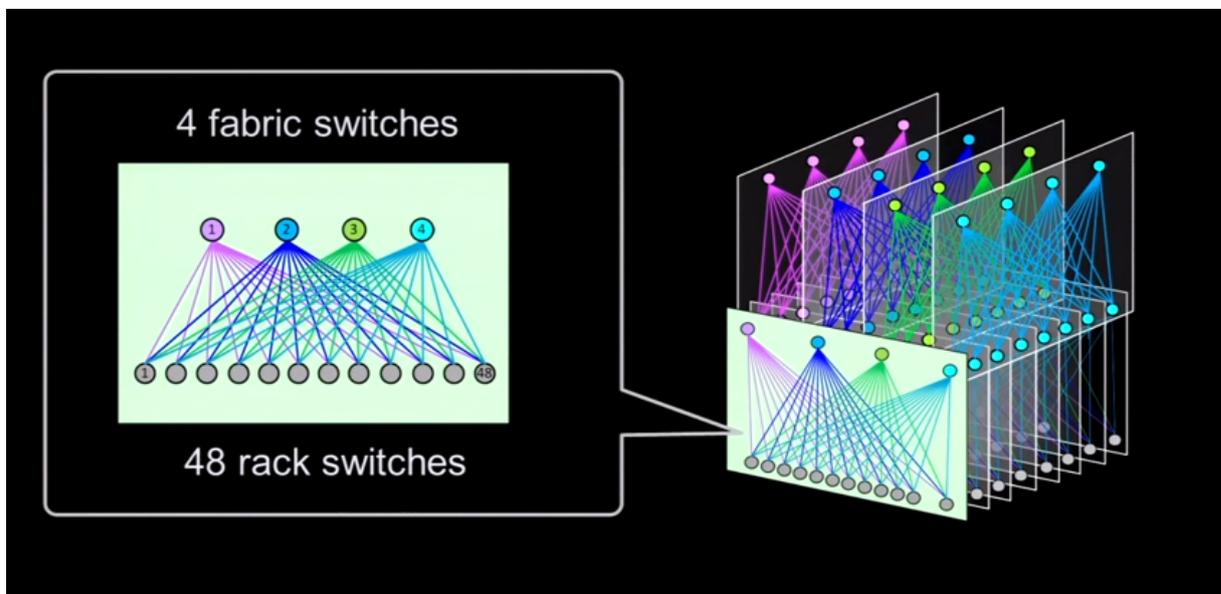


Figure 3.2 Illustration of a Server Pod which has only 48 server racks (Top-of-rack Switches) [5].

One major difference between cluster-based design and fabric design is that although contiguous rack switches are placed physically in the cluster, rack switches within the same pod are built with no hard physical limitations within a datacentre.

3.3.2 Traffic pattern inside Facebook's datacentre network

This study involves applying real Facebook traffic patterns, which are based upon the data collected from the fabric network design (a state-of-the-art) of the Facebook data centre, see Figure 3.1. The Facebook data was sourced from the Github repository. The accessed data consists of raw Facebook traffic patterns from three of Facebook's clusters running different applicants – *Frontend*, *Hadoop* and *Database* over a period of 24 hours. Nevertheless, the proposed method also would be valid to a diverse range of network data sets and scenarios.

We focused on an individual “Database cluster” within their massive collection of network traffic. The Facebook network flow data includes both the inbound and outbound traffic (packet traces). Each record in the Facebook data comprises “sampled flow” information, which contains packet header information as well as some metadata such as locality and packet length. When extracted, the *Database* cluster’s data size exceeds 50 gigabytes and contains around 316 million packets.

We visualise the Facebook data set with a node-link diagram to explore what is on a network through a visualisation tool in Power BI. The network (see Figure 3.3) is shown as a graph with servers being nodes and the packet length being flows on edges. This graph represents how the server-level data communicate with each other within the *Database* cluster over a period of 24 hours. The simplicity and beauty of node-link diagrams turn into clutter and confusion as the graph of the Facebook real-world dataset is so massive. When the standard force-directed method was applied, the resulting graph resulted in a hairball drawing.

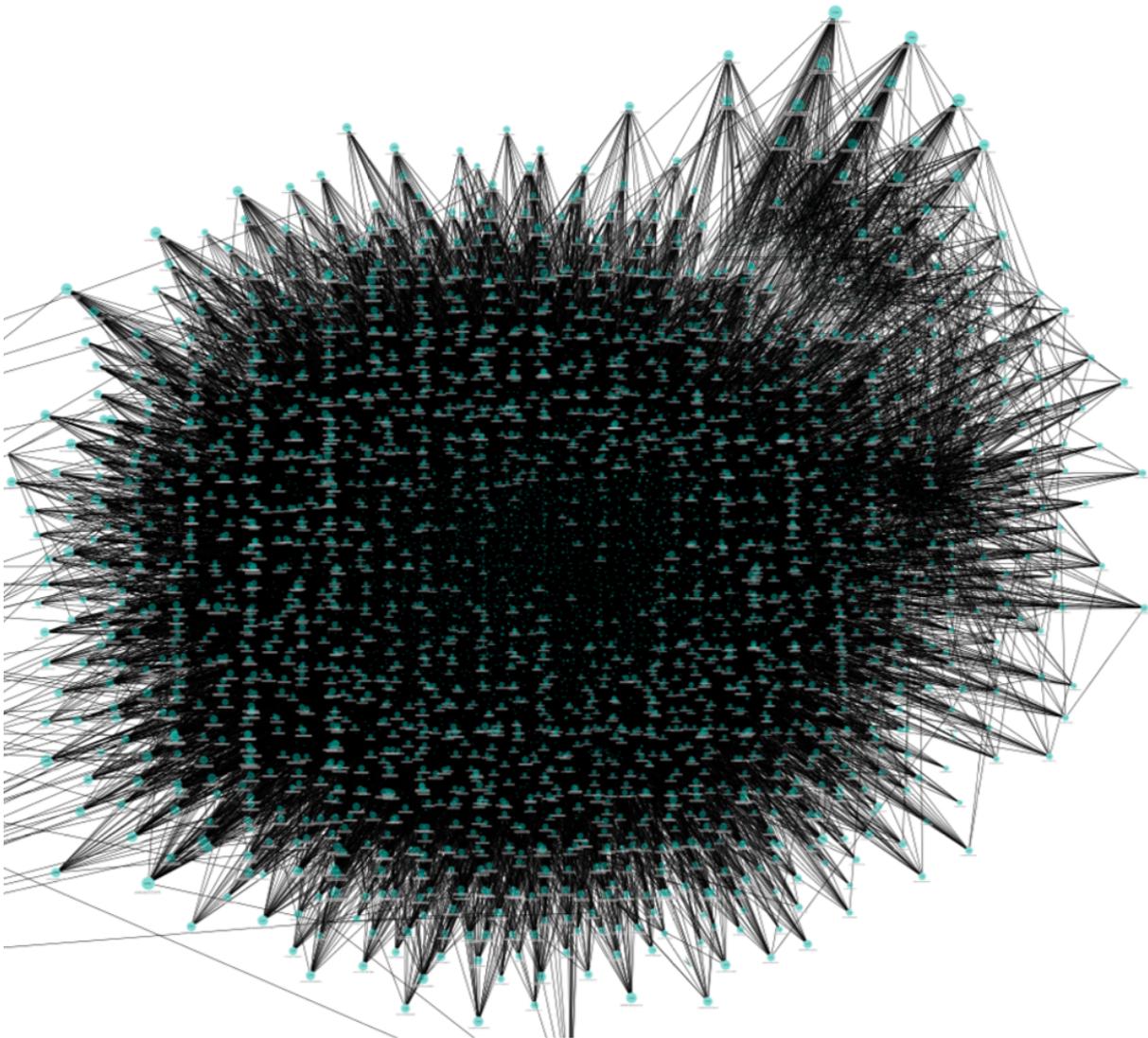


Figure 3.3 “force-directed” layout algorithm is used within the visualisation tool for representing 245349 distinct source IPs (server machine) and 116152 distinct destination IPs. The graph was zoomed in on a particular section of a larger graph. (Source: Author’s own)

Network traffic visualisation can be employed to detect abnormally high network activity and promptly locate server machines that send or receive a substantial number of packets. The utilisation of conventional techniques to depict complex and large graphs, as demonstrated in Figure 3.3, results in a more challenging task for humans to interpret and comprehend the visual representation of such graphs. Consequently, this leads to a failure to identify intriguing patterns that are not easily detectable using existing graph visualisation methods.

We present an algorithm to simplify the input graph prior to visualisation, with the aim of producing simplified subgraphs. This enhances the graph's readability by lowering the displayed edges on the graph and prioritising "interesting" relationships before visualisation. In our research, the interestingness is determined by edge variance.

3.4 Concept of the Variance in edge weights

In this thesis, to reduce the complexity, the primary emphasis is placed on comprehending the data connected with the edges and nodes, rather than the graph structure and connectivity aspects, such as graph clustering, which are of lesser significance. Thus, we focus on the simplification technique that does not prioritise the demonstration of the overall graph structure.

Summarisation involves revealing "interesting" nodes and edges. A common strategy is to represent only the relevant information, much as the irrelevant information is hidden. However, the definition of "interesting" is subjective and usually requires domain knowledge and user preferences to be taken into account. For example, in some applications, the most interesting information may be the edges with the highest weight, while in other applications, the most interesting information may be the edges that form the largest connected component. In our research, our metric focuses on **edge variance**. This research holds importance in enhancing the comprehension of the intrinsic aspects of network elements rather than solely focusing on their observable properties or the overall data structure.

3.4.1 Edge Variance

Variance and standard deviation are fundamental statistical measures that enable us to comprehend how flat or insightful the data is. Both variance and standard deviation provide a quantitative measure of uncertainty in data. These measures are essential for understanding the degree of uncertainty and variability in data. Thus, to show uncertainty in the output, the uncertainty in the user input has to be quantified.

The variance is a statistical measure used to quantify the degree of spread or dispersion within a given data set. The variance for a population, denoted by σ^2 is defined as follows:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} \quad (3.1)$$

Where x_i represents each value in the dataset, μ is the mean of the data, and ‘N’ is the total number of observations in the dataset. However, for a sample variance, the variance σ^2 is slightly different, using ‘N-1’ (degrees of freedom) instead of ‘N’.

The emphasis of our metric is **edge weight variance**. Thus, we study edge variance to have an idea of the edge weight homogeneity or heterogeneity present in the graph-based data. The edge variance concept aims to provide end-users with visual representations that incorporate and reflect the variance of the information to aid in informed decision-making and more efficient data analysis. The variance in edge weights, denoted by $Var(E)$, which is a measure of how much the weights of the edges between a pair of distinct nodes deviate from their mean value, is defined to be :

$$Var(E) = \frac{\sum_{i=1}^n (w_i - \mu)^2}{N} \quad (3.2)$$

Where w_i refers to edge weight, μ is the mean of the measure values, and ‘N’ refers to the total number of edges between a pair of distinct vertices.

3.4.2 Applying Edge Variance Concept to Facebook Traffic Patterns

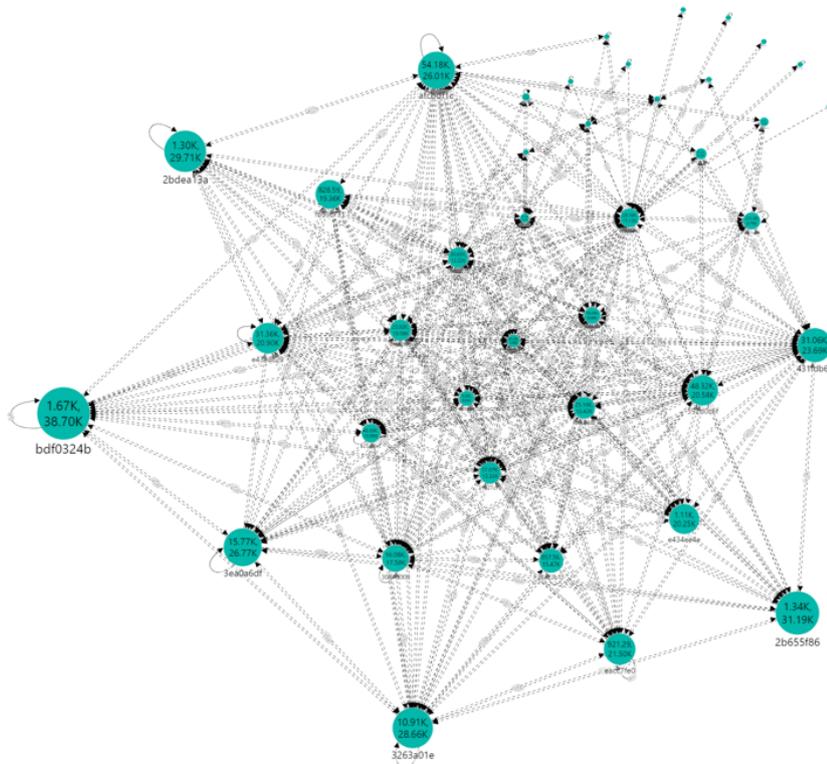


Figure 3.4 The network was shown as a graph with server racks being nodes and the variance of bandwidth being flows on edges. (Source: Author’s own)

The network depicted in Figure 3.4 illustrates a graph with server racks being nodes and the variance of bandwidth being flows on edges. This represents how the rack-level data

communicate with each other, and bi-directional links, as in real networks, illustrate the variance of the bandwidth with the direction. The mathematical formula, defined in 3.2, was used to quantify the variance of bandwidth from source rack to destination rack traffic per link.

As a result, in a graph, edges with high variance in their weights may indicate that the weight of the edge is not stable, and the edge may not be reliable. In contrast, edges with low variance in their weights may indicate that the weight of the edge is stable and the edge is likely to be reliable. For example, variance in edge weights can be helpful in identifying patterns in the graph, such as edges that represent outliers or deviations, which may be important for certain applications, such as cyber security, in which a large variance could potentially signal the presence of an attack or vulnerability. In a nutshell, measuring and comprehension of the variance in edge weights can contribute to a more complete comprehension of network dynamics.

In a summary graph, when the number of edges to be shown is limited, the proposed approach not only decreases the size of the graph but also discerns edges that contain the most useful information from an information-theoretic perspective.

3.5 Information Theory

Information theory can be defined as “the science of quantification, coding and communication of information” [190]. It is commonly agreed that the field of information theory was introduced, and most of its underlying problems were solved by Claude Shannon in his pioneering work of “A Mathematical Theory of Communication” in 1948 [166]. At its core, information theory focuses on the quantification of the amount of information contained in a message, known as the entropy of the message [167].

Various information-theoretic measures, such as Shannon’s information measures (entropy, conditional entropy, mutual information [44]) for discrete and continuous random variables, and relative entropy (or Kullback-Leibler divergence) [46], have been applied to solve various problems [44].

3.5.1 Shannon Entropy

Entropy, a concept rooted in statistical physics, is a fascinating and complex notion that encompasses a multitude of definitions and a wide range of contexts [70]. Shannon directed his attention towards the issue of effectively encoding the information intended for transmission by a sender, ultimately formulating the concept of information entropy as a metric for quantifying the level of uncertainty present inside a given message.

The concept of entropy, first used by Shannon to study communication channels, can be used to quantify the degree of uncertainty, randomness, or disorder in a graph [131]. This is because both communication channels and graphs can be seen as sources of information with a certain probability distribution, and entropy provides a measure of the amount of uncertainty or information content in that distribution [166].

A singular and definitive definition of graph entropy does not exist [70]. It is plausible that no definitive or exemplary solution exists, as the effectiveness of a particular approach in one context

may not be applicable or advantageous in another. Hence, an alternative conceptualisation of probability could result in distinct formulations for the graph entropies.

The concept of entropy in a graph could be understood as quantifying the graph's structural information (e.g., numbers of vertices, edges, node degrees, and distances), hence functioning as a metric for complexity measure. Different graph entropies have been widely employed to characterise the structural properties of networks across diverse disciplines [48, 49]. However, it is important to note that our attention is not directed towards particular structural features of a graph. In other words, our objective is not to assess the level of structural complexity to identify the specific type of structural information being measured by these metrics. Here, the term "entropy of a graph" refers to an information-theoretic measure that is applied to a graph with a discrete probability distribution on its edge weight set.

We first discuss some notions regarding the *Shannon information content* and *Shannon entropy*. The notion that clarifies how to calculate the information content of a random variable is formally characterised as follows:

Let X be a triple containing x, A_X, Pr_X parameters, in which the outcome of x is the discrete random variable, that holds a finite number of values $A_X = \{x_1, x_2, \dots, x_m\}$, having probabilities $Pr_X = \{p_1, p_2, \dots, p_m\}$ with a probability $p(x_i)$. And, let p be a *probability mass function* of X and is defined as $p(x_i) = P(X = x_i)$, $p(x_i) \geq 0$. Our definitions for Shannon information content employ only discrete probability distributions over finite sets A_X . The following two lemmas enable us to work with a discrete probability distribution $p(x_i)$.

Lemma 3.5.1.

$$\sum_{x_i \in A_X} p(x = x_i) = 1$$

Lemma 3.5.2. *Entropy is additive for independent random variables. If X and Y are two independent random variables, then the entropy of the joint random variable $H(X, Y)$ is equal to:*

$$H(X, Y) = H(X) + H(Y) \text{ iff } P(x, y) = P(x)P(y)$$

The *information content* of an outcome, x , whose probability is $p(x_i)$, is defined as follows:

$$h(x = x_i) \equiv \log_2 \frac{1}{p(x_i)} \quad (3.3)$$

The *mean entropy* of any set of probabilities is formalised to define the average Shannon information content of an outcome [166]:

$$H(X) := \sum_{x_i \in A_X} p(x_i) \log_2 \frac{1}{p(x_i)} \quad (3.4)$$

Throughout the work, Logarithm \log function is taken to the base two, being \log_2 to represent the information content. Thus, the entropy is expressed and measured in bits.

The binary entropy, as depicted in Figure 3.5, refers to the measure uncertainty associated with a random variable x that has an alphabet consisting of two elements, namely $\{x_1, x_2\}$.

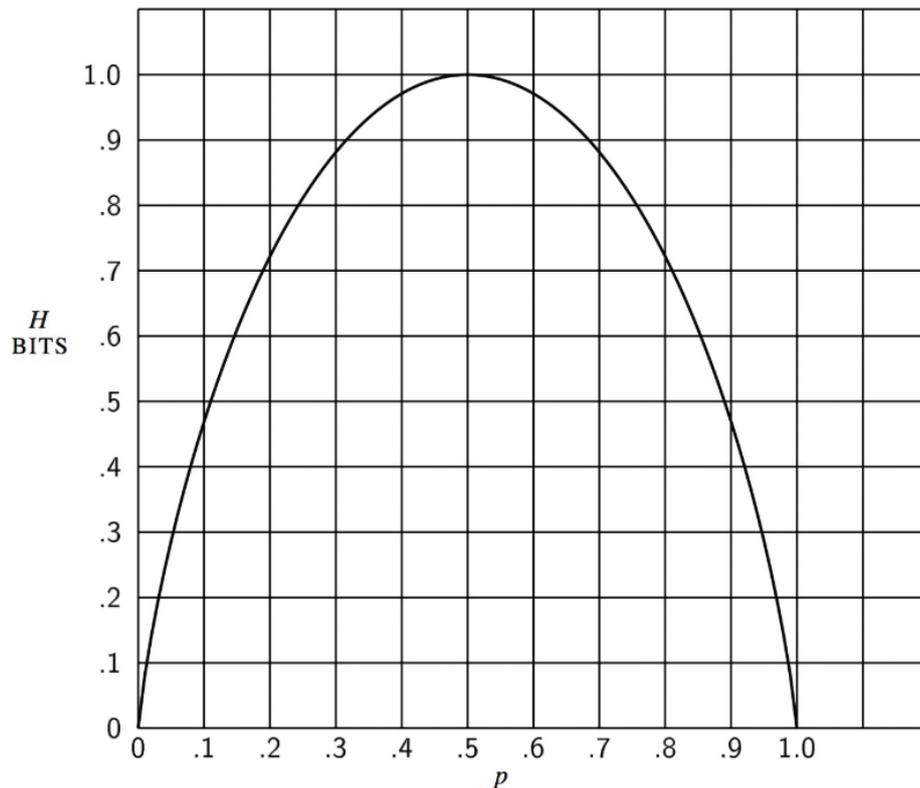


Figure 3.5 Plot of binary entropy used by Shannon in [166].

The probability distribution of this random variable is denoted by $\{p, 1 - p\}$. The formula for calculating the binary entropy is given by

$$H(X) = -p \log p - (1 - p) \log(1 - p) \quad (3.5)$$

A discrete random variable is applicable in describing the outcome of a fair coin flip. In this case, the alphabet represents the possible outcome, which is $A_X = \{head, tail\}$. The probability distribution indicates that each option has an equal likelihood of occurring, with a probability $P_X = \{1/2, 1/2\}$. The binary entropy reaches its maximum value of 1 bit at a time when $p = 1/2$, because the results are independent. Thus, the entropy of a fair coin flip is given by $H(X) = -(1/2) \log(1/2) - (1/2) \log(1/2) = \log 2 = 1$ bit. However, when the coin deviates from being fair, the situation becomes intriguing. When the coin is not equitable, such as a single event predominates, as in skewed probability distribution, there is a reduced level of surprise/uncertainty, resulting in a lower entropy within the distribution.

If we transition from a distribution with skewed probabilities to one with equal probabilities, it can be anticipated that the entropy will exhibit a pattern of initially being low and subsequently increasing. More specifically, the entropy will start at its minimum value of 0 for events that have an impossibility and certainty (with probabilities of 0 and 1, respectively) and will reach its maximum value of 1 for events with equal probabilities. Consequently, in scenarios where no event holds dominance over another, such as when the probability distribution is equal or approximately equal, we anticipate a maximum entropy, see Figure 3.5.

A further exemplification can be provided by considering the scenario of rolling a fair six-sided dice. In the context of a fair dice toss, denoted by the alphabet $A_X = \{1, 2, 3, 4, 5, 6\}$ and the probability distribution assigns equal probabilities of $1/6$ to each outcome, defined as $P_X = \{1/6, 1/6, 1/6, 1/6, 1/6, 1/6\}$. The entropy of the random variable X , denoted by $H(X)$, is calculated as $\log_6 6$, which is approximately equal to 2.58 bits. To summarise, the entropy of probability distribution is bounded between 0 and $\log_2(N)$, where N represents the overall count of events within the distribution.

- When the probability distribution is skewed, entropy will be low.
- When the probability distribution is balanced, entropy will be high.

The key to successfully incorporating the notion of entropy into visualisation problems relies on properly specifying the random variable X and constructing the probability function $p(x)$. Consider a graph $G = (V, E)$, where V represents the set of vertices and E represents a set of edges. The edge variance between a pair of nodes within a given edge set forms a discrete probability distribution by treating the graph as a discrete random variable. The entropy of the graph G is calculated by computing the probability of each distinct edge in the graph. The mathematical expression for the entropy, denoted by $H(G)$, is as follows:

$$H(G) = \sum_{i=1}^n Pr(e_i) I(e_i) \quad \text{or} \quad - \sum_{i=1}^n Pr(e_i) \log_2 Pr(e_i) \quad (3.6)$$

Which probability distribution on graphs should we choose? The maximum entropy principle is the crucial determinant in finding the answer. A graph with a higher entropy value would signify a higher degree of randomness or uncertainty in the distribution of edges, while a graph with a lower entropy value would imply a lower degree of randomness or uncertainty in the distribution of edges. In this study, we propose a summary algorithm that forms a subgraph that seeks to maximise its entropy based on information theory [63, 166]. The entropy of the edge weights (variance) distribution of the summarised graph is maximised amongst all possible aggregations of a given edge size in terms of Shannon entropy.

3.6 Entropy-Based Graph Summarisation Algorithm

3.6.1 Overview

The proposed summarisation method in this chapter is inspired by *Maximum Entropy Summary Trees*, studied by Karloff and Shirley [99]. We propose a method similar to their work in that both methods summarise the relational data while maximising Shannon's entropy. Our approach deals with non-hierarchical graph-based data to determine which information is more relevant to be displayed while they focus on node-weighted rooted trees (hierarchical graphs).

According to Karloff and Shirley [99], the optimal selection from a set of summary trees with a predetermined number of nodes is the one that maximises the entropy of the probability

distribution linked to the summary tree. They define the concept of entropy, where the normalised weight of each tree node gives the probability density function.

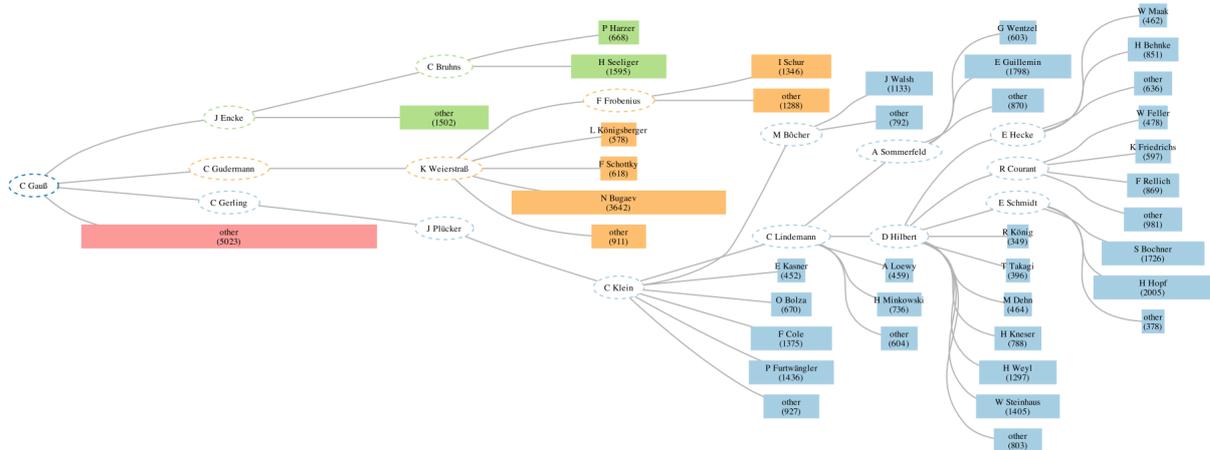


Figure 3.6 A screenshot from the study of Karloff and Shirley showing the 56-node summary tree of the math genealogy, utilising the maximum entropy approach. The colours of nodes are determined by their depth-1 ancestor, while the sizes of nodes are proportional to their weights in the summary tree. [99]

The above Figure 3.6 depicts an instance of a maximum entropy summary tree within the context of a mathematical genealogy tree [99]. Certain nodes are depicted in their original structure, taking the form of ellipses. Conversely, other nodes are utilised to represent subtrees and are depicted as rectangular shapes featuring a name and a weight. Additionally, there are nodes that symbolise the aggregation of sibling subtrees, taking the form of rectangular shapes, but with a *other* label and the associated weight.

We introduce how to summarise (non-hierarchical) graph-based data by extracting the maximum information in readable and informative forms of the original graph to end-users. This study focuses on (directed or undirected) weighted multigraphs with non-negative edge weights. In this study, the Shannon entropy function is adopted as the primary metric in order to measure information of non-hierarchical graphs from an information-theoretic perspective. The core idea of the proposed algorithm is to merge several “less interesting” edges into one edge called *superedge* and uncovers high-level “more interesting” edge patterns, resulting in a subgraph having a set of unique edges collected during the procedure. Our main concern is maximum information retrieval; consequently, the proposed method proposes a rule for selecting a distribution which maximises Shannon’s entropy of the summarised graph of a given length of the edge.

3.6.2 Implementation

Our algorithm works on a weighted directed or undirected multigraph \mathbf{G} with each edge \mathbf{e} having an associated real-valued (non-negative) number $\mathbf{w}(\mathbf{e})$, called its **weight**. The inclusion of weights on a graph is a valuable source of supplementary data that should be taken into account during the analysis. This study proposes an algorithm for finding a summary graph based on the specified edge length. The algorithm generates output in the JSON-formatted data, consisting of

multiple summary graphs that begin with two edges and increase to a specified number of edges. This feature enables end-users to observe the inclusion of edges during each iteration through interactive means.

Finding distinct edges

The algorithm starts by computing a new graph $G' = (e, w, Pr)$ from the input multigraph G , which is the first step in the decomposition of the multigraph into a graph G' having a set of multiple distinct edges with their weights (variance) as well as its probabilities. The newly constructed graph G' has at least three parameters: e represents a set of distinct edges, w represents the variance of the weight of the distinct edges, and Pr represents the edge probability distribution. The edge probability distribution Pr is calculated by dividing each edge's weight by the sum of all edge weights, which forms a probability list of distinct edges. The important steps of the procedure are described as follows:

1. Let $G = (V, E, W)$ be a non-empty, finite directed or undirected, weighted non-hierarchical multigraph where weights are nonnegative and assume $V = \{v_1, v_2, \dots, v_n\}$ and $E = \{e_1, e_2, \dots, e_m\}$.
2. Form an $G' = (e, w, Pr)$, has at least three parameters and denotes the number of distinct edges by $d = |e|$ with $d \geq 3$.
 - Let $e_i = u, v$ be a distinct edge ($e_i \subseteq E$).
 - Let each edge e_i has weight w_i , denoting the variance of the distinct edge $e \in E$. The weight w_i of the distinct edge $e_i = u, v$ is computed as the variance of the corresponding weights over all edges between a vertex u and a vertex v .
 - Pr represents an edge probability distribution, and each of these distinct edges holds a non-negative probability from 0 to 1.
3. Let $Pr_i := \frac{w_i}{\sum_{i=1}^d w_i}$, denotes the probability distribution of each distinct edges.

Forming summary graphs

The second step is forming summary graphs (S) that capture the most "interesting" edges from the graph G' having distinct edges, and viewers can choose the n desired number of edges to be shown. This means end-users wish to see a summary graph comprising $(n-1)$ selected distinct edges and one *superedge*. The "superedge" results from the rest of the edges merging into an edge. The probability of the "superedge" is set to 1 minus the sum of the probabilities of the selected edges. This step is important as it helps to reduce the number of edges in the summary graph S while maintaining the overall probability distribution of the graph G' . By applying Shannon's Eq. (3.6) with the probability distribution, we obtain a numerical value that indexes the maximum information content. Figure 3.7 provides a flowchart showing the step-by-step process by which the algorithm functions on graph G' .

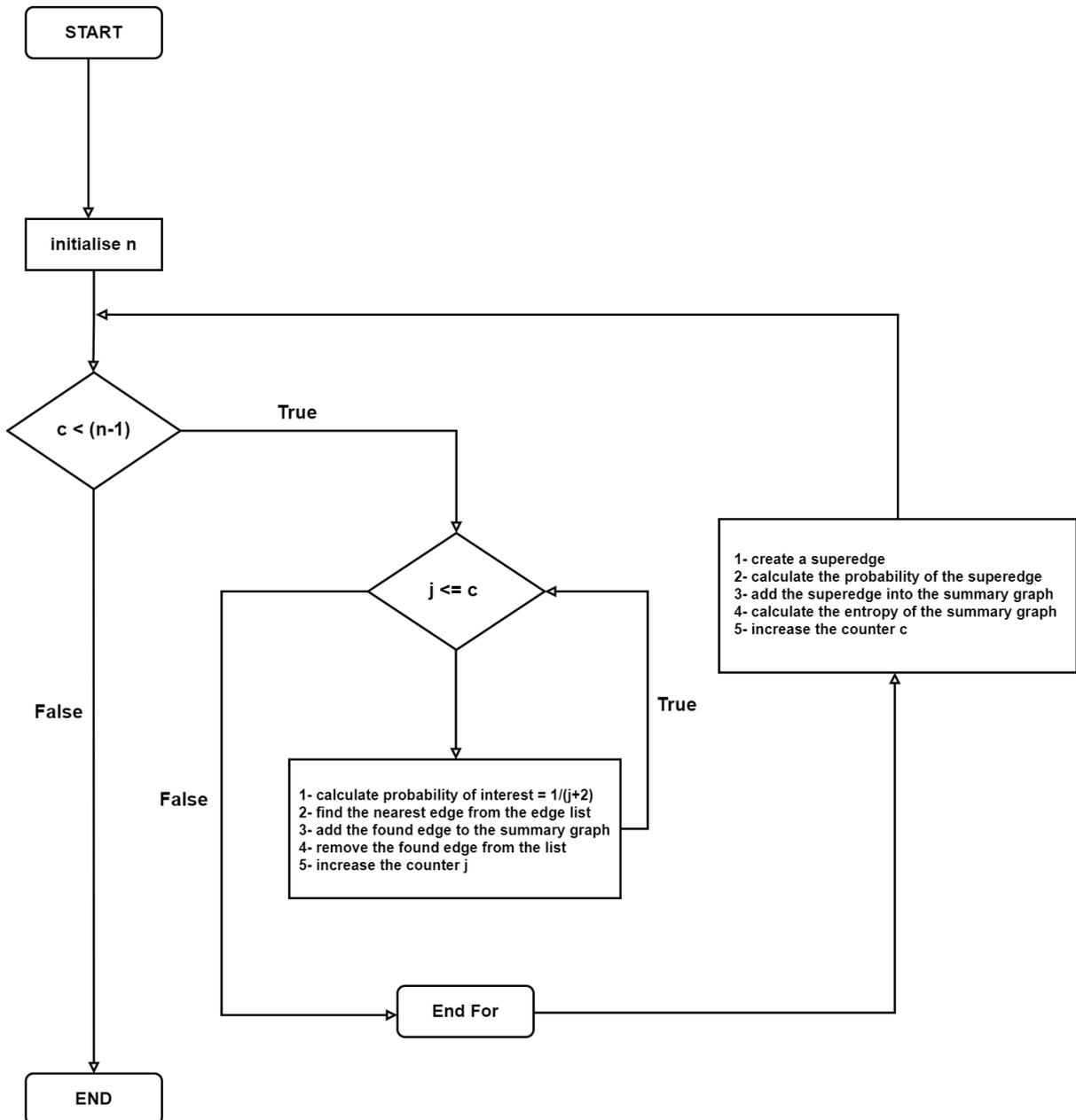


Figure 3.7 The flowchart outlines the formation of separate edges, commencing from a minimum of two edges and continues until the desired number, as specified by the user, is reached.

Shannon entropy function applies for a distinct edge probability distribution. The concept of “equal probability, maximum entropy” lies at the core of the construction of probability function algorithm design. The idea behind our algorithm is to adopt this supporting argument so that the entropy of the summarised graph is maximised amongst all possible combinations of a given edge size. It should be emphasised that entropy is convex and maximised when $Pr(i)$ values are uniform. When the probability of selected edges is uniform, the Shannon entropy will be higher, indicating that the resulting summary graph captures most of the information of the original graph in the summary graph. The detailed whole graph summarisation algorithm is described in Algorithm 1 as follows:

Algorithm 1: Graph Summarisation Algorithm for a directed or undirected Graphs

```

Data: Graph  $G' = (V, E, W)$ ,  $n =$  desired number of edges in the summary graph
Result:  $S = (V, E)$ 
 $c \leftarrow 0$ ; /* iteration counter */
 $S = (V, E) \leftarrow \emptyset$ ; /* empty summary graph */
while  $c < n - 1$  do
   $G' = (e, w, Pr)$ ; /* list of distinct edges */
  for  $j = 0$ ;  $j \leq c$ ;  $j++$  do
     $p_i \leftarrow \frac{1}{j+2}$ ; /* Calculate probability of interest */
     $nearest\_edge \leftarrow find\_nearest\_edge(p_i, Pr)$ ;
     $S_c.append(nearest\_edge)$ ; /* Add edge to summary graph */
     $remove\_edge(nearest\_edge, G')$ ; /* Remove edge from  $G'$  */
  end
   $superedge \leftarrow merge\_edges(G')$ ;
   $superedge.Pr \leftarrow 1 - Sum(Pr)$ ;
   $S_c.append(superedge)$ ;
  calculate entropy  $H(S_c)$ ; /* based on the Algorithm 3.6 */
   $c++$ ; /* increase the iteration counter */
end

```

For example, summarising a graph with two edges ensures that it contains one unique edge selected from all available edges and one *superedge*. *Superedge* consists of two distinct nodes: *meta-source* and *meta-destination*. The *meta-source* node merges all other source nodes, while the *meta-destination* node merges all other destination nodes. Consider a scenario where an edge with a probability of 0.1 is chosen, forming a *superedge* with a probability of 0.9, leading to an entropy value of 0.47. Conversely, choosing an edge that is closest or equal to 0.5 will yield the maximum entropy for the specified summary graph. The maximum entropy approach searches for edge distributions having properties we desire in the most informative way.

The generated JSON-formatted array includes the probabilities of edge variance, source nodes, destination nodes, and entropy of the graph for each summarised graph, starting from edge number 2 up to the edge number defined by the user. The array also contains statistical data for the edges, including mean, variance, minimum, and maximum values. See an example of Figure 3.8 of the returned JSON-formatted 4-edged summarised graph.



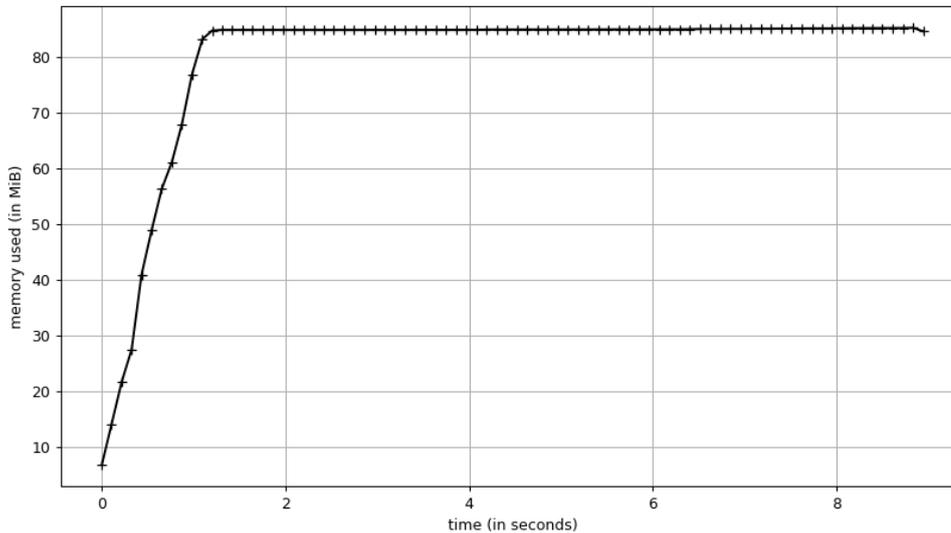
Figure 3.8 A screenshot of the 4-edged graph summary, which the proposed algorithm generates results in JSON (Source: Author’s own figure)

3.7 Testing with Facebook network data

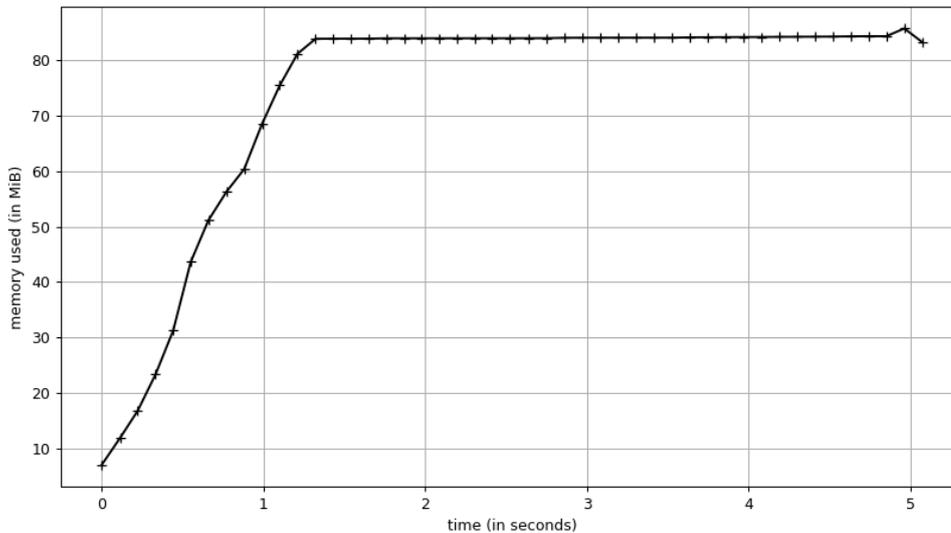
We tested our approach to simplify the analysis of network traffic inside of Facebook servers [5]. The real dataset used for testing consists of Facebook traffic patterns from a weighted directed network. We measured the execution time (in seconds) and memory consumption of the proposed algorithm on a laptop. The test was executed using a 64-bit Windows operation system with an AMD Ryzen 7 2.9 GHz CPU and 8 GB memory. We ran the algorithm around ten times to find a 100-edge summary graph, and used the mean execution time as a result.

The efficiency of summarisation algorithms is a crucial concern due to the abundance of data on real networks. The computational complexity of an algorithm refers to the estimation of the resources, such as how much time or memory it uses, that are necessary for the algorithm to execute a given task. Typically, the scalability of such demands is indicated by their correlation with the scale of the system under investigation. When considering a graph, its size is commonly denoted by the count of vertices n and/or the count of edges m .

In complexity theory, linear growth is denoted as $O(N)$, signifying that the runtime of a process increases according to the number of elements involved. As seen in Figure 3.9, a graph with up to 1500 distinct edges finishes in under 10 seconds and requires about 5 seconds for a graph with 1000 distinct edges. Resulting in the algorithm preserving linear execution time. As a result, the approach exhibits favourable scalability as the graph size increases.



(a) Test result displays the execution time (in seconds) and memory usage (in MiB) of the summary algorithm applied to a dataset of 1500 distinct edges.



(b) Test result displays the execution time (in seconds) and memory usage (in MiB) of the summary algorithm applied to a dataset of 1000 distinct edges.

Figure 3.9 Figures present the relationship between execution time and memory usage of the algorithm when processing Facebook network data with differing complexities—1500 edges in (a) and 1000 edges in (b).

Note that the edge list is sorted in the previous step when we find the distinct edges from the multigraph. In conclusion, the algorithm works linearly if the edges are sorted. However, if the list needs sorting, the best sorting method will be $O(N \log N)$.

Overall. In general, our approach enables identifying patterns and trends within network traffic data, such as which servers handle the most traffic and which traffic is the most important

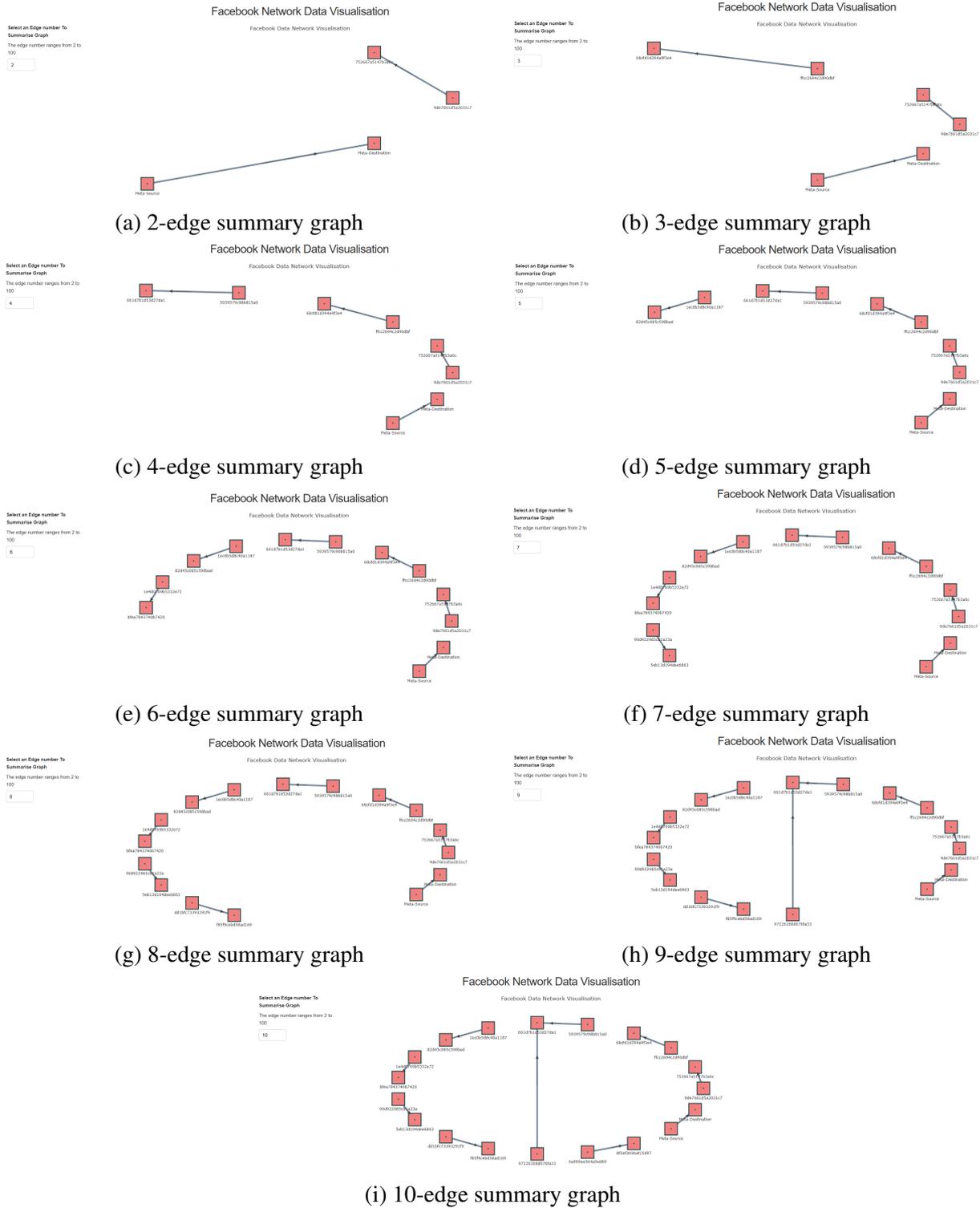


Figure 3.11 For a given number of $n = 10$ edges, the resulting summary graph $S = (V, E)$ consists of the collection of sets of distinct edges comprising $(n - 1)$ selected distinct edges and one *superedge*. Each generated summary subgraph exhibits maximum information given the same number of edges from an information theoretical point of view.

As an illustration, the summarisation algorithm could be used to provide simplified summaries of the transportation networks. The edge weights may represent factors such as passenger volume or traffic patterns. For example, traffic on a given road or ridership between two locations varies at different times of day. Variations in ridership or traffic patterns throughout the day are incorporated into the algorithm, potentially serving as a key variable of interest. The proposed algorithm may help identify roads with significant fluctuations in traffic count, indicating high traffic variability. Additionally, it might identify routes with significant fluctuations in passenger counts.

3.8 Limitations

One limitation of Shannon entropy is that it is only applicable for discrete random variables. In other words, it can only be used to calculate the amount of information in a finite set of discrete outcomes or events. If the input multigraph has continuous variables or edges with continuous weights, this algorithm would not be able to use Shannon entropy to calculate the amount of information captured in the summary graph. In such cases, other measures of information or uncertainty, such as differential entropy or Rényi entropy, should be used instead. It should be noted that while these methods can help overcome the limitation of Shannon entropy, they also come with their own limitations and trade-offs. Therefore, choosing the right method to use would depend on the specific characteristics of the input multigraph and the goal of the analysis.

The function “*find_nearest_edge*” in the Algorithm 1 is designed to identify the probability value that is closest to the desired *probability of interest*. This is done by searching through the list that contains all the probabilities associated with each distinct edge in G' and subsequently selecting the nearest match from the list. However, in the case of multiple identical values for the nearest edge, the process was intentionally designated to follow a first-come, first-served methodology.

The proposed algorithm does not consider that some nodes may be more interesting than others, and it does not consider the weight of the nodes; these factors may lead to the summary graph missing some important information that was represented by the original graph. So, it can be said that the algorithm tries to capture the most interesting edges based on maximum entropy. Still, considering the variability of nodes, it is not guaranteed that the resulting summary graph will have the most information from the original graph.

3.9 Future work

As we focused on the concept of the variance in edge weights, particularly edge uncertainty, the resultant summary allows the user to focus on more “interesting” edges, defined in terms of edge variance. As future work, we could implement a threshold value for edge variance in the graph summarisation approach. Instead of taking into account all edges in the graph G' , this threshold would eliminate edges with variance lower than a predetermined value. This methodology can improve efficiency by prioritising edges with more significant variations in characteristics within

the graph, which may reduce computational complexity. Further exploration is required to determine suitable threshold selection methods and assess the practical advantages across diverse applications.

In graph analysis, the significance of edges may extend beyond mere variability, and certain edges, albeit displaying low variability, can still assume crucial roles. The proposed algorithm holds promise for future extensions, especially when considering incorporating a weighting factor for edges. Currently, the method assigns equal weight to all edges in relation to their variance, which may not fully capture the significance of specific edges. Edges of high importance, but having low variance, might not be readily observed within the present context, hence potentially hiding key insights. Incorporating a weighting factor into the algorithm allows for a more sophisticated approach, as it prioritises edges based on their variance and weighting factor. This, in turn, may facilitate a graph summarising process that is more context-aware and produces a more precise outcome. This improvement aligns with the algorithm's adaptability and offers the potential for a more comprehensive and insightful analysis of relational datasets.

In situations where understanding both nodes and edges is crucial for decision-making and analysis, integrating the concept of node variability could be a potential future improvement to the graph summarisation algorithm, which adeptly addresses edge variability. To effectively implement this adjustment, the algorithm could include a variance calculation for each node, measuring the variability in data attributes. For example, in a telecommunications network, while edge variability might capture the fluctuation in data flow between nodes, node variability can reflect the differences in data handling capacities or traffic frequencies of individual nodes. By integrating this dual approach, the algorithm would distinguish interesting edge patterns and ensure that such interesting nodes are also included in the network summary. As another example, in healthcare networks, nodes representing hospitals might have weights based on patient intake and treatment success rates, while edges might have patient referral volumes. By adapting the algorithm to prioritise nodes with higher variability, such as those hospitals with variable patient intake, the summary graph would more accurately reflect critical aspects of the healthcare network. This enhanced approach could improve the algorithm's applicability and effectiveness in real-world scenarios.

3.10 Summary

This chapter introduces an entropy-based graph summarisation algorithm to highlight "interesting" patterns within large, complex networks. This approach fundamentally diverges from traditional methods, which typically prioritise structural simplification or sheer reduction in size without considering the underlying information content. Unlike traditional methods that may overlook the interplay of edge characteristics within the graph, this approach utilises Shannon entropy and uses edge weight variance as a metric to ensure that the most informative edges are retained in the summarised output. By focusing on high-variance edges, the method ensures that the essential characteristics of the original graph are not only preserved but are also made more discernible.

This technique marks a significant advancement over existing summarisation methods by providing a more data-driven, information-theoretic foundation for graph reduction, which enhances our understanding of what features of the graph are most critical to retain during the summarisation process. This method can provide a more effective means of analysing large graphs, which is particularly valuable in domains such as social network analysis, bioinformatics, and network security, where understanding the most interesting connections can lead to significant discoveries.

The effectiveness of the approach is showcased through JSON-formatted output and visual representations of each summarised graph, as illustrated in Figure 3.11 and Figure 3.8, respectively. However, the incorporation of supplementary edge variables, including primary value (e.g., mean value) and secondary value (e.g., variance), into the visual representation of the summary graph will greatly assist users in understanding the properties of the subset of edges selected from the original graph. Additionally, visualising these features is most likely useful for expedited graph analysis, the rapid comprehension of crucial information, the immediate recognition of data patterns, and other related benefits. These motivations inspired the work discussed in the next chapter, which will introduce a novel edge design that enhances the process of mapping bivariate data encompassing both quantitative and qualitative features in the context of node-link diagrams.

Chapter 4

Introduction Vizent Edges

Contents

4.1	Introduction	78
4.2	Motivation	78
4.3	Related Works	79
4.4	Glyph-based Visual Design	80
4.5	Introduction To Vizent Edge Design	81
4.5.1	Visual Entropy (Vizent) Edge Design	83
4.5.2	Null Case Representation	83
4.5.3	Experimental Process of the Edge Glyph	84
4.6	Results	87
4.6.1	Effect Size	88
4.7	Summary	90

Overview

This chapter contains materials previously published by Akbulut et al. [1], and it introduces the concept of the novel *Vizent* edge, designed based on established design principles and inspired by the previous study proposed by Holliman et al. [80]. Vizent edge design was designed for representing ordinal categorical or numerical values on a discrete scale. We employed the empirical study, and the intention is to establish if the reported order of the glyphs followed the frequency complexity of the pattern. The empirical results of the Vizent edge design showed that the edge glyph in variations of pattern frequency can convey the ordered data. We also conclude that the linear Vizent edge glyphs are highly distinguishable in order by pattern complexity in pairwise testing. The main contributions of this study are:

- Presenting the concept of the novel Vizent edge design;
- Empirically validating (in collaboration with Lucy McLaughlin) that different edge glyphs have a perceived order through pairwise testing.

4.1 Introduction

Node-link diagrams, the most prominent visual encoding idiom, represent network data due to their simplicity and intuitiveness [207]. Network data attributes can be directly labelled over the links (or edges) and nodes, such as textual value, or mapped to them through encoding channels. Shape, colour, size, texture, and value are a few examples of such visual channels adapted from Bertin [23]. Numerous methods exist for visually representing node attributes. However, few studies exist for effectively encoding edge data attributes on edges in node-link diagrams [135]. Moreover, there are only a limited number of visual variables to visually display bivariate (two attributes) information on edges simultaneously [154]. Even though there are chances to combine these variables, their practical range of values is also somewhat restricted [154]. For example, visual variables that affect the overall visibility of a line mark may not always be robust in the inclusion of another visual variable (e.g., fuzziness and width), resulting in interference with accurate visual perception [71].

4.2 Motivation

Node-link visualisations enable users to analyse, evaluate, and explore networks while also facilitating a deeper understanding of interesting aspects of the data using the proper visual channels [50]. In addition, it is worth noting that node-link diagrams offer enhanced readability and familiarity, presenting the benefit of a relatively shorter learning curve for effective utilisation compared to alternative types of representation, such as matrix representations, for graphs of smaller sizes [67].

Communication networks are primarily concerned with the information flow amongst people or devices such as server machines [135], and what matters most is the size of the flows

and how they change over time for decision-makers. Network traffic visualisations can help analysts detect unusually large amounts of packets [171]. It is also interesting to monitor the variability of this traffic to recognise network anomalies. It is crucial that decision-makers should be aware of the uncertainty that may be present; otherwise, the visual analysis might not draw accurate and precise conclusions and comprehend the nuances of relational datasets, making visual analysis imperfect and error-prone [97]. However, one fundamental problem arises regarding how to provide additional uncertainty values alongside the data value while maintaining comprehension [29].

4.3 Related Works

In this section, we summarise studies on edge visualisation for representing uncertainty information. We limit the scope to static graph representations that can be printed on paper or visualised on an electronic display.

Uncertainty has emerged as an active research area within the field of visualisation due to requirements on dependability and interpretability in data analysis. Visualisation of uncertainty is regarded as one of the five major technical challenges in visualisation [117]. Various visualisation methodologies have been used to visualise uncertainty across various disciplines [97]. However, the visualisation of uncertainty in graphs is an important and emerging subject of research that has gained much interest in recent years [206].

The values of graph properties with uncertainty are visualised through two distinct visual channels, and the edge is a significant visual primitive that encodes graph attributes. The visual appearance of the edges can be modified to depict uncertainty regarding edge existence, location, or attribute values [196]. However, node-link diagrams have not been employed to visualise edge attributes until recently [163]. Generally, graph practitioners focus on the edge shape (e.g., curved, tapered) [82, 210], the performance of graph layouts [147, 145] clutter reduction techniques (edge bundling) [128, 81]. For example, Holten et al. [82] evaluated edge shape representation with different approaches such as tapered and glyph patterns for edge directionality.

MacEachren et al. [122] proposed the theory of visual semiotics of uncertainty. They evaluated the intuitiveness of various visual variables, such as fuzziness and brightness, in representing uncertainty through the participants' judgements.

Boukhelifa et al. [27] presented sketchiness as a visual variable for displaying univariate (one attribute) uncertainty value on edges and compared its efficiency to other visual variables such as dashes, blurring, and grayscale. Additionally, Schwank et al. [163] evaluated four visual approaches for representing uncertainty on edges: dashes, waves, stripes, and blurring. They concluded that dashes and blurring appear to be particularly effective amongst other options at representing edge uncertainty [27, 163]. However, blurring and transparency have portability issues if the monitors they are used on are not calibrated to have the same output luminance given the same input value. Adjustable gamma curves can affect this directly.

Guo et al. [71] experimentally evaluated several visual encodings used for representing bivariate edge attributes, main value, and uncertainty, to determine which pair of visual encodings

performs better in node-link diagrams. It was found that fuzziness strongly interfered with the perception of an edge when paired with width, and that width and fuzziness had an adverse effect on accuracy.

Bae et al. [13] also examined how different visual encodings affect the interpretation of bivariate variables on directed edges. They found that visualisations utilising numbers and visualisations with width and brightness performed more accurately than those with hue and granularity. However, their evaluations in [71, 13] did not address the challenges that needed participants to search for dual visual attributes on edges simultaneously. Also, they experimented with one fixed graph size of 18 nodes and 25 edges.

Several studies have already addressed the depiction of univariate uncertainty in node and edge attributes. However, incorporating uncertainty into graph attributes remains an outstanding issue. Existing works displaying uncertainty about edge attributes need to be strengthened to visualise two or more edge attributes effectively [196].

4.4 Glyph-based Visual Design

Glyph-based visualisation [25] is a widely used visual design form in which variables of different data types, both categorical and numerical, are encoded by pre-defined visual metaphors known as glyphs. Bertin [23] introduced retinal variables that can be employed to depict information, offering options for encoding data in glyph designs, such as shape, size, or value. With this in mind, we present the concept of the novel Vizent edge design that employs a combination of colour and variations in stripe pattern, which we call edge glyph. The definition of a glyph that we adhere to, as proposed by Borgo et al. [25], is described as follows: “A glyph is a small independent visual object that depicts attributes of a data record; glyphs are discretely placed in a display space; and glyphs are a type of visual sign but differ from other types of signs such as icons, indices and symbols.”

The prevalence of these visualisations in modern life is due to their effective utilisation of the human ability to interpret abstract and metaphorical depictions, hence facilitating rapid recognition and comprehension of information. Glyphs are commonly employed in diverse contexts, including node-link diagrams, treemaps, and geographic maps, owing to their space-efficient graphical appearance. Glyphs encode one or more data values of their appearance (also called visual channels).

Chen and Floridi [35] devised a taxonomy method to categorise more than 30 visual channels. Among the numerous options available for visual channels, the prevalent approach in multivariate visualisation is utilising multiple visual channels. These are classified into the following channel categories: *Geometric channels*, such as orientation and shape; *Optical channels*, such as hue and texture; *Topological and Relational channels*; and, lastly, *Semantic channels*, as can be seen in Figure 4.1 below. Visual channels for conveying information can be effectively utilised within the domain of glyph design.

There have been abundant glyph design guidelines proposed in the literature. The guidelines concern various levels of glyph design: *variable encoding*, *inter-channel interaction*, and *holistic glyph design*. At the variable encoding, Bertin [23] introduced a framework for classifying

Geometric Channels	Optical Channels	Topological and Relational Channels	Semantic Channels
<ul style="list-style-type: none"> ● size / length / width / depth / area / volume ● orientation / slope ● angle ● shape ● curvature ● smoothness 	<ul style="list-style-type: none"> ● intensity / brightness ● colour / hue / saturation ● opacity / transparency ● texture (partly geometric) ● line styles (partly geometric) ● focus / blur / fading ● shading and lighting effects ● shadow ● depth (implicit / explicit cues) ● implicit motion / motion blur ● explicit motion / animation / flicker 	<ul style="list-style-type: none"> ● spatial location ● connection ● node / internal node / terminator ● intersection / overlap ● depth ordering / partial occlusion ● closure / containment ● distance / density 	<ul style="list-style-type: none"> ● number ● text ● symbol / ideogram ● sign / icon / logo / glyph / pictogram ● isotype

Figure 4.1 A Various Range of Visual Channels [35]. (Source: [25])

semantic relevance, which may be used to assess the appropriateness of various channels for representing specific categories of information. Cleveland and McGill [40] identified the accuracy of human perception in various visual variables and gave recommendations on the choices of visual channels for different tasks and purposes of visualisation. At the levels of inter-channel interactions and holistic glyph design, guidelines were proposed for the *typedness*, *channel capacity*, *visual orderability*, *integration and separability* of channels, *searchability*, *learnability of glyph designs*, *attention balance*, and *focus and context* [38].

The foundations and design guidelines of glyphs were examined by Borgo et al. [25]. It has been suggested that glyphs can attract more attention and elicit higher levels of cognitive activity during visualisation compared to other types of visual design. In their study, Maguire et al. [124] proposed a series of design principles for visual encoding. These principles were derived from the psychology literature on perception and visual search areas. The guidelines encompass considerations related to *semantic relevance*, *channel composition*, *pop-out effects* (also known as visual pre-attentiveness), and *visual hierarchy*. As a result, they suggested the effective design of glyphs should incorporate multiple retinal variables that are separable and do not conflict with each other in terms of channel composition.

4.5 Introduction To Vizent Edge Design

Design is a crucial factor in node-link visualisations. In graph design, studies on edge visualisation have shown that edge design influences graph reading [197, 135]. The initial design choice that must be made when addressing the visualisation of graphs pertains to the visual representa-

tions of entities and relationships, namely their mapping to two-dimensional geometric shapes, also known as planar primitives. Given the limitations of manipulating mapping information in a static context, it becomes necessary to encode it in a manner that enhances understanding for the observer. A well-designed visual representation of graph attributes in the node-link graph can facilitate effective and timely visual search and pattern identification.

Glyph-based visual design often encounters challenges when visually representing multiple data properties. One notable problem pertains to perception, namely the level of ease associated with comprehending and accurately interpreting visualisations of this nature. The act of encoding data into various visual attributes of a glyph in a simplistic manner does not inherently result in a well-designed outcome. In fact, it has the potential to perplex viewers or even result in incorrect interpretations in the most unfavourable scenarios. The size of glyphs also imposes limitations on their design when compared to a whole visualisation. For instance, smaller glyphs have different design considerations compared to larger glyphs. This constraint affects the number of variables that can be physically encoded and presented on a screen without compromising the information's integrity.

The core of information visualisation comprises two features: a mark, in our case, which is a planar primitive and a visual channel, which determines the appearance of marks [23, 133]. Using these two aspects of visual encoding should comply with the following two principles: *expressiveness* and *effectiveness*. It is essential to consider the possible interactions of visual channels when using more than one in visual encoding. Using independent separable channels results in practical visuals instead of inevitably joined integral channels. For example, a glyph is likely composed of a set of visual variables, and hence, the composition of these channels may exert an influence on the manner in which individual channels are apprehended. Therefore, we integrate suitable visual channels in a pre-attentive way to achieve effective visual channel composition.

Holliman et al. [80] defined the *visual entropy* of shape to be a measure of its complexity as perceived by a human viewer. They provided a novel set of glyphs for displaying the uncertainty of a measure alongside its mean value in both 2D and 3D visualisation environments, as seen in Figure 4.2. Motivated by this, we propose the notion of visual entropy (Vizent) edges to address the research gap and advance the field of graph visualisations for representing bivariate variables on edges in node-link diagrams.

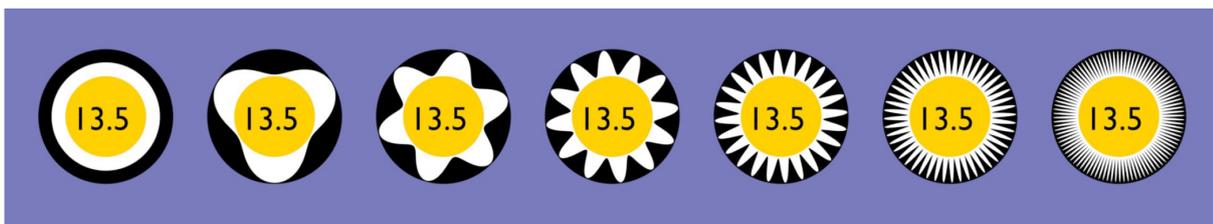


Figure 4.2 Holliman et al. [80] proposed a novel set of visual entropy glyphs. The value increases from left to right proportionally to the complexity (frequency) of the shape.

4.5.1 Visual Entropy (Vizent) Edge Design

Vizent edges represent ordinal categorical or numerical interval values on a discrete scale. Fig. 4.3 illustrates an example presenting how bivariate edge variables are mapped to the Vizent design. For the two-variable challenge, our design solution is static and 2D. This design employs a combination of colour and central stripe patterns on edges in node-link diagrams. Edge colour, related to data value, can be set on a predefined colour scale. In contrast, the second value is encoded into a stripe pattern as an edge glyph and displayed in the centre of the edges. The method consists of a carefully designed 7-step sequence of stripe patterns that shows an uncertainty value as a second value and a colour sequence. As with the original glyphs [80], the pattern frequency is doubled to make them highly distinguishable.

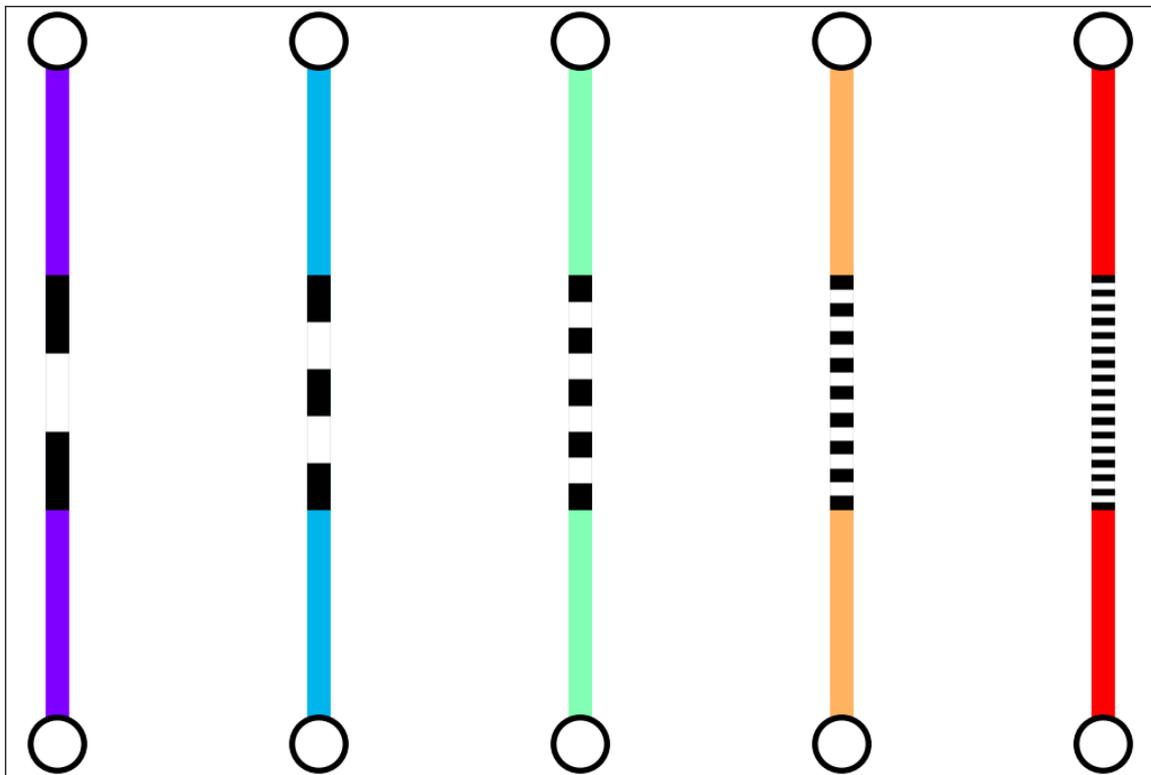


Figure 4.3 An example of a new node-link visual model. Visual Entropy (Vizent) edge design combining colour and stripe pattern.

4.5.2 Null Case Representation

The Vizent edge design depicted in Figure 4.3 can visually represent the two-variable challenge, such as displaying data value and a second value. However, it is also important to develop a glyph design that accurately conveys the scenario in which there is an absence of data concerning the second value. To be exemplified, this pertains to situations wherein we possess data value (such as a variable's mean value) that requires representation, although we lack any knowledge regarding its corresponding level of uncertainty. Figure 4.4 depicts the Vizent edge representing the scenario where uncertainty measures and/or data values are absent.



Figure 4.4 This edge glyph design is developed for cases where the dataset exhibits instances of missing uncertainty value.

4.5.3 Experimental Process of the Edge Glyph

Establishing an order in which objects are perceived becomes readily achievable when a visual channel is ordered. A key idea in our approach to constructing Vizent edge glyphs is that they can be ordered visually.

The primary purpose of the experiment is to empirically examine the hypothesis that a perceived rank ordering exists among the visual entropy edge glyphs depicted in Figure 4.5 below. We employed a “two-alternative forced choice” (2AFC) [74] methodology involving the comparison of edge glyph image pairs to assess the existence of a rank ordering between the patterns. Each participant was presented with all possible combinations of paired permutations from the set of glyphs, excluding pairs with identical pattern frequency. Subsequently, the participants were required to select one of the presented pairs.

To confirm the perceptual effectiveness of the Vizent edge glyphs, we adapted our experimental design for the original Vizent glyphs to the linear glyphs. This asked participants: “Each image in the pair represents a value using a repeated pattern of bars. If more complex repeated patterns of bars represent higher values, then please choose which image represents the higher value to you.” The intention is to establish if the reported order of the glyphs followed the frequency complexity of the pattern.

PsychoPy toolbox [142] was used to implement the presentation of the stimulus [142]. The participants exclusively utilised the keyboard as a means of response in order to minimise any potential delays caused by cursor movement. In addition to capturing the time taken to input their

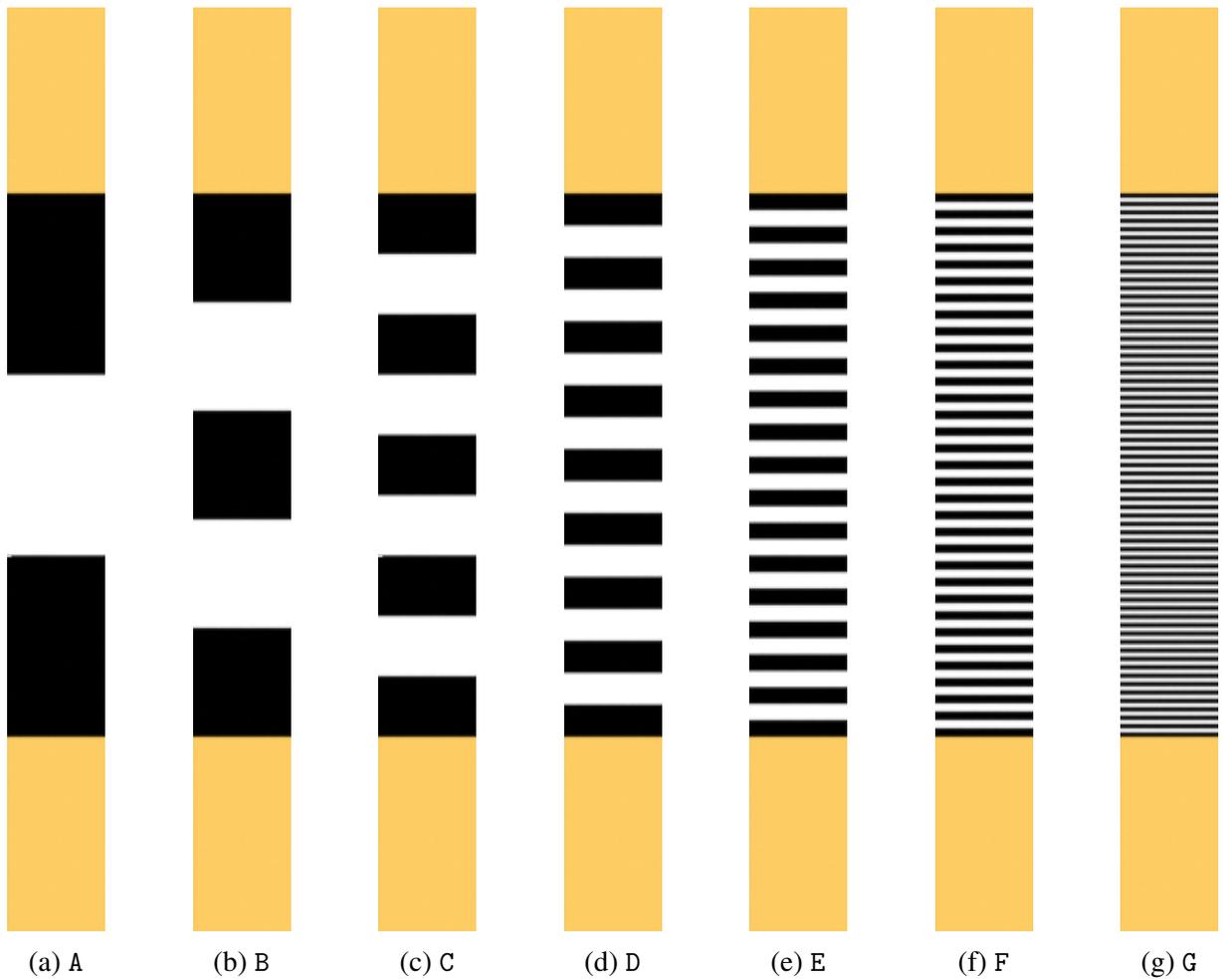


Figure 4.5 The novel design of the Visual Entropy (Vizent) edge glyphs. In the experiment, every pair of a set of seven edge glyphs were shown for comparison.

responses, the participants' answers for each pair comparison were recorded by registering the left or right arrow key-press. Each participant was exposed to a different and random sequence of pairs, which was determined using PsychoPy's random number generator. The instructions each participant read before the trial are shown in Figure 4.6 below.

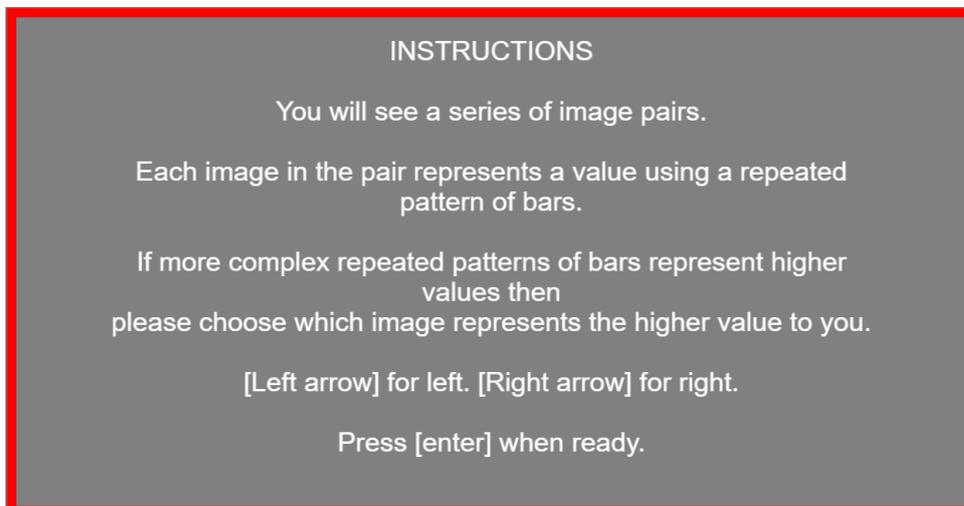


Figure 4.6 The figure shows the instructions that were presented to each participant prior to the commencement of the trial.

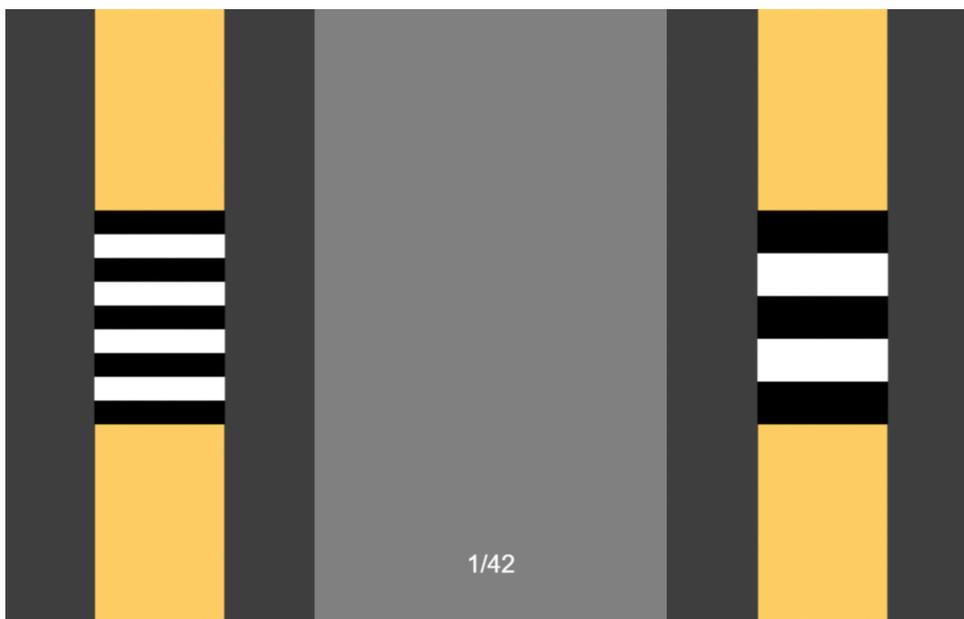


Figure 4.7 An example of a pairwise comparison image used in the trial.

We specifically obtained very limited personal details, inquiring about participants' visual acuity status (normal or corrected to normal) and whether they were conscious of any colour impairment in their visual capabilities. Although it is unlikely that colour deficiency had an impact on the frequency complexity of the edge glyph comparison experiment, we made note of this possibility to consider its potential influence on the saliency of the coloured edge design. No data was recorded on demographic differences, such as age, gender, and relevant expertise factors, as we had no expectation there would be differences. Newcastle University provided an

ethical review and approval of the experiment. Participants permitted their data to be used and began the trial.

We recruited participants on Prolific [140], a crowdsourcing online platform with a research focus. A total of 29 participants were shown every pair of a set of seven of the edge glyphs assigned as A to G, as shown in Fig. 4.5, using PschoPy [142] and Prolific [140] to present the experiment. The participants were provided with a total of 42 (" $7 \times 6 = 42$ ") pairwise permutations of the edge glyph in Fig 4.5, which included reversed order image pairs. An example of such a pair presented in the experiment can be seen in Figure 4.7 above.

4.6 Results

We ensured the integrity of the data by implementing statistical outlier detection. We specifically applied z-score analysis to identify participants whose response times were unusually long or short (z-scores greater than 3 or less than -3). The results were screened for outliers. Following applying a threshold of three standard deviations from the mean to the results, two of the 29 participants were excluded, leaving 27 participants' data for further analysis.

All participants reported normal or corrected to normal vision. The independent variable is each glyph type (A, B, C, D, E, F, G), and the dependent variable is the proportion of correct choices made regarding the order of each glyph in the pairwise comparisons with the remaining six other edge glyphs. With $n = 27$ participants, the total number of trials per glyph is " $27 \times 6 = 162$ ". This value represents the maximum achievable correct score for each glyph.

Table 4.1 The table illustrates the edge glyph pairwise order comparisons alongside the outcomes of the exact binomial test for each individual glyph.

Glyph	Correct	Trials	p-value	probability	CI low	CI high
A	161	162	0.0056	0.994	0.966	1.
B	162	162	< 0.001	1.	0.977	1.
C	161	162	0.0056	0.994	0.966	1.
D	161	162	0.0056	0.994	0.966	1.
E	162	162	< 0.001	1.	0.977	1.
F	162	162	< 0.001	1.	0.977	1.
G	161	162	0.0056	0.994	0.966	1.

The accuracy rates of the participants, those in agreement with the prediction that the higher frequency would be rated as more complex, were analysed using a G-Test for Goodness of Fit, as recommended in [180]. This tested the null hypothesis that there was no difference in the response accuracy scores between the glyphs. Each glyph was expected to have $1/7$ of the total correct decisions. The alternative hypothesis is that there is a statistically significant difference in the proportions of correct answers for some of the glyphs. The result was ($G = 0.0106$,

$p < 0.999$, $df = 6$) and since $p \gg \alpha = 0.05$ we could not discard the null hypothesis, and there is no evidence of a significant difference in the accuracy scores between the glyphs.

Given that the outcomes represent a count of categorical selections, we apply an exact binomial test, as suggested in [127], for each glyph tested whether the accuracy was higher than the threshold performance. We set the threshold to be at the standard psychometric level of 75%. Therefore, the null hypothesis is that performance should be no different to the 75% level. For all glyphs, the performance was significantly higher than this ($p < 0.001$) and was also higher than a 95% threshold ($p < 0.01$). Therefore, participant accuracy is significantly different (better) than the null hypothesis. The outcomes of the binomial tests are shown in Table 4.1 and depicted in Figure 4.8.

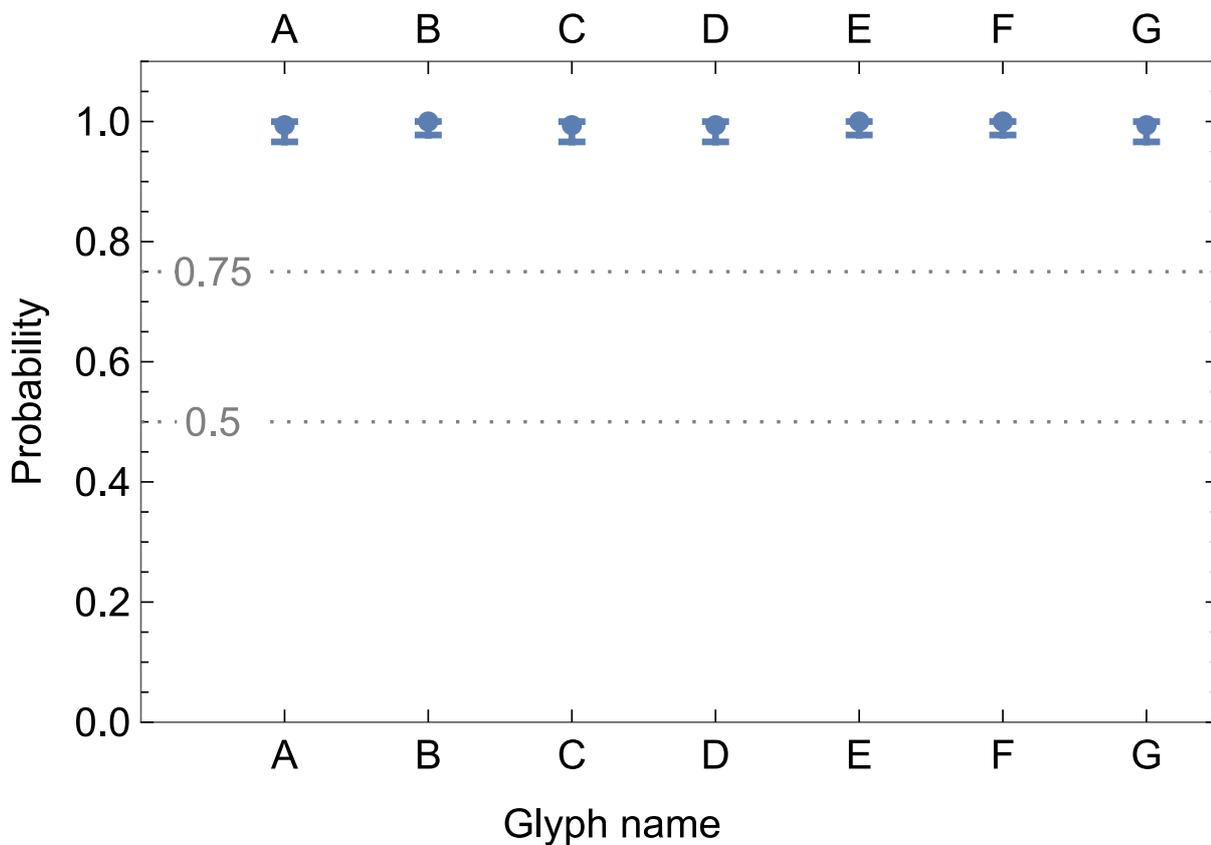


Figure 4.8 The results show the probability of correct response with 95% confidence intervals for each glyph, ($n=27$).

We collected response time data for each comparison and found no overall difference between the glyphs in a one-way ANOVA ($F = 0.267$, $df = 6$, $p < 0.963$), see Fig. 4.9. Therefore, no post hoc tests were justified for response time.

4.6.1 Effect Size

Effect sizes and confidence intervals (CI) provide an indication of the magnitude of the observed differences across experimental conditions. This process facilitates a deeper comprehension of the underlying factors supporting our hypothesis and concepts, enabling us to enhance and fine-tune our ideas and forecasts in subsequent research endeavours [32]. One methodology

Response Time(s)

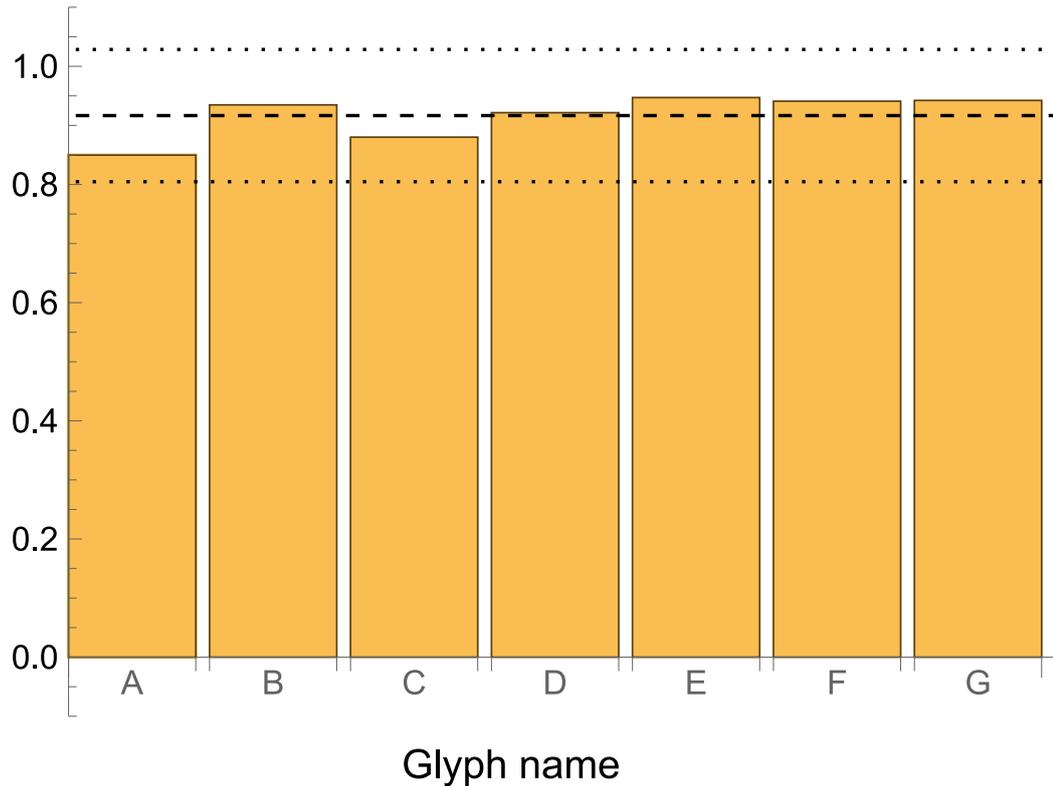


Figure 4.9 Mean response time for each glyph, (n=27).

commonly used to estimate effect size for binomial tests, as suggested in [143], is the calculation of Cohen's g [41]. This technique can be employed to estimate the effect size specifically for probabilities compared to a chance result of 50% correct responses.

As illustrated in Table 4.2, Cohen's g is categorised into four distinct classifications. Regarding effect size, a value of g below 0.05 is regarded as *negligible*, whereas a value below 0.15 is categorised as *small*. Furthermore, a value less than 0.25 is classified as *medium*, and any value equal to or greater than 0.25 is considered as *large*. Cohen's g effect sizes are calculated for each glyph as shown in Table 4.3. These results suggest that the effect size for all the glyphs is *large* ($g > 0.25$).

Table 4.2 Interpretation of Cohen's g effect size.

range	effect size
$0.00 \leq g < 0.05$	Negligible
$0.05 \leq g < 0.15$	Small
$0.15 \leq g < 0.25$	Medium
$g \geq 0.25$	Large

The practical significance of effect size is contingent upon the specific situation, as elaborated extensively in [32]. Based on the considerable probability of users accurately selecting the

Table 4.3 Effect Size Estimate Using Cohen's g For Each Of The Edge Glyph Binomial Tests

glyph	probability	g	effect
A	0.994	0.494	Large
B	1.	0.5	Large
C	0.994	0.494	Large
D	0.994	0.494	Large
E	1.	0.5	Large
F	1.	0.5	Large
G	0.994	0.494	Large

appropriate sequence, it is justifiable to infer that the practical effect size is strong. In fact, almost every user judges the order of the edge glyphs according to their suggested order.

4.7 Summary

Bertin [23] introduced a framework consisting of four semantic categories to determine the appropriateness of various visual channels, known as retinal variables, for conveying specific types of information. These semantic criteria are *associative*, *selective*, *ordered* and *quantitative*. Bertin [23] stated that the texture channel, meaning patterns such these, could be ordered. However, he did not present any evidence to support the ordered perception of the texture. Our results verify Bertin's proposed hypothesis that the texture variation in linear representation can convey the ordered data. We also conclude that the linear Vizent edge glyphs are highly distinguishable in order by complexity in pairwise testing and show no significant variation in response time by complexity. This supports our choice to compare the Vizent edge representation in detail with previous proposals for displaying bivariate edge data to determine if they are as good or better than existing alternatives.

Chapter 5

A novel node-link visual model – visual entropy (Vizent) graph

Contents

5.1	Introduction	92
5.2	Experiment 1: Edge Performance Experiment	93
5.2.1	Usability Study of Bivariate Network Visualisation Approaches	93
5.2.2	Control Group of Visual Encodings	94
5.2.3	Tasks	98
5.2.4	Hypotheses	99
5.2.5	Experimental Procedure	99
5.2.6	Edge Performance Experiment Results	100
5.2.7	Graph Size	105
5.2.8	Task Type	109
5.3	Experiment 2: Node and Edge Performance Experiment	110
5.3.1	Hypotheses	113
5.3.2	Experimental Results	113
5.3.3	Participants' comments on the experiment	115
5.4	Discussion and Conclusion	115
5.5	Limitations	117
5.6	Future Works	117

Overview

This chapter presents and evaluates our proposed node-link visual model through user studies. Most of the work in this chapter was previously published in [1]. Here, we present two experiments that evaluated the performance of our proposed method compared to existing methods. The first experiment discusses (see section 5.2) the methodology and the results of a task-based usability study for comparing four different bivariate network visualisation approaches. The second experiment (see section 5.3) evaluated the usability of the Vizent edge design with the Vizent glyphs proposed by Holliman et al. [80]. Lastly, we discuss the results and limits of the research study and a roadmap for future research in this area.

In summary, the main contributions of this study are:

- Presenting a novel node-link visual model – visual entropy (Vizent) graph – to effectively represent both primary and secondary values, such as uncertainty, on the edges simultaneously.
- Performing two task-based usability studies to demonstrate the efficiency and effectiveness of our approach for visualising bivariate networks using static node-link diagrams.
- Comparing the Vizent design against three visual encodings selected from the literature on various graphs ranging in complexity from 5 to 25 edges for three different tasks.

5.1 Introduction

Networks (or graphs) have become ubiquitous in diverse fields, such as social, communication, and transportation networks [197]. Given networks increasing size and complexity, network visualisation goals are shifting to the challenge of readability and the rapid extraction of valuable insights from networks [30]. However, a challenge in visualising network data is developing an accurate mental model of what is happening in a network so that appropriate actions can be taken quickly.

Visual network analysis has received growing interest due to data availability across all application domains. Many tasks involve examining the characteristics of the nodes and edges that can be subject to uncertainty. This uncertainty can take several forms, depending on the context. It can be quantified with a single scalar value and described as variability, probability, confidence, or other measures. Introducing uncertainty, even as a simple scale scalar value, could complicate the representation. The difficulty of utilising existing methods, increasing visual complexity, and the absence of effective visualisation techniques complicate the visualisation of uncertainty. Interactive visualisation approaches and user interaction might address these concerns [97]; however, interactivity is not a solution in many situations, such as print and books, and is only applicable in limited circumstances. Incorporating uncertainty information into existing visual channels explicitly as an additional channel in statically depicted graphs might present a readability issue.

The main goal of this study is to examine and enhance the limitations of the existing visual encodings by developing a novel approach of a node-link visual model — visual entropy (Vizent)

graph. All research problems motivated us to provide a solution to be effective and simple to interpret on (static) node-link diagrams. We test the hypothesis that Vizent performs equally or better against three visual encodings selected from the literature based on the performance of the predefined user tasks. Therefore, we formulated the following research question:

1. Does the Vizent design perform equally well or better against the three visual encodings for all tasks in terms of response time and accuracy?

In most cases, investigating the limitations of scalability may not be the main focus of experimental studies. Instead, researchers may ignore scalability by selecting a size range within which they are comfortable conducting their study [211]. Unlike literature, we are also interested in evaluating whether the performance of each visual encoding varies with graph complexity and task complexity. This prompted us to perform a user study including graphs of varying sizes and tasks of varying degrees of difficulty. As a result, we examined two secondary research questions:

1. Does the varying number of graph edges significantly influence the performance of each visual encoding?
2. Does the type of task (*Single* or *Dual* target visual search) significantly impact the performance of each visual encoding?

5.2 Experiment 1: Edge Performance Experiment

This section describes the control group of visual encodings, tasks, hypotheses tested, the experimental procedure that we employed to conduct the user study, and results.

5.2.1 Usability Study of Bivariate Network Visualisation Approaches

Vizent edge design and three different visual encodings were employed for comparison in our first experiment, see Figure 5.1. This experiment involves synthetically generated network traffic patterns with a node-link diagram. Each edge has two values labeled as “Network traffic” and “Variability”. Network traffic was presented as the primary value, and the uncertainty information was presented under a variability level as a secondary value to highlight the use and potential of our proposed design for analyzing communication networks. Nevertheless, the proposed node-link visual model also would be valid for a diverse range of networks.

Network traffic ranges in value between 20 and 80 in increments of 20, and the variability ranges in value between 1 and 3. Four encoding levels were assigned to network traffic, while three were assigned to its variability. Each edge represents one of the combinations of a network traffic level and its variability level. In the experiment, each displayed graph has just one edge with the desired value, which all the participants are required to locate. For the remaining edges, the values of each edge were always selected from a discrete uniform distribution; values were drawn from 20,40,60,80 for the network traffic, and values were drawn from 1,2,3 for its variability.

5.2.2 Control Group of Visual Encodings

This section summarises three pairs of visual encodings selected for comparison. We replicated the following visual encodings from Guo et al. [71] and Bae et al. [13] that were found to be effective: Width–Lightness, Numerical values, and Saturation–Transparency. The following abbreviations are therefore used throughout the remainder of this study to refer to the visual encodings: Width–Lightness = *Wid–Lig*, Numerical values = *Num*, Saturation–Transparency = *Sat–Tra*.

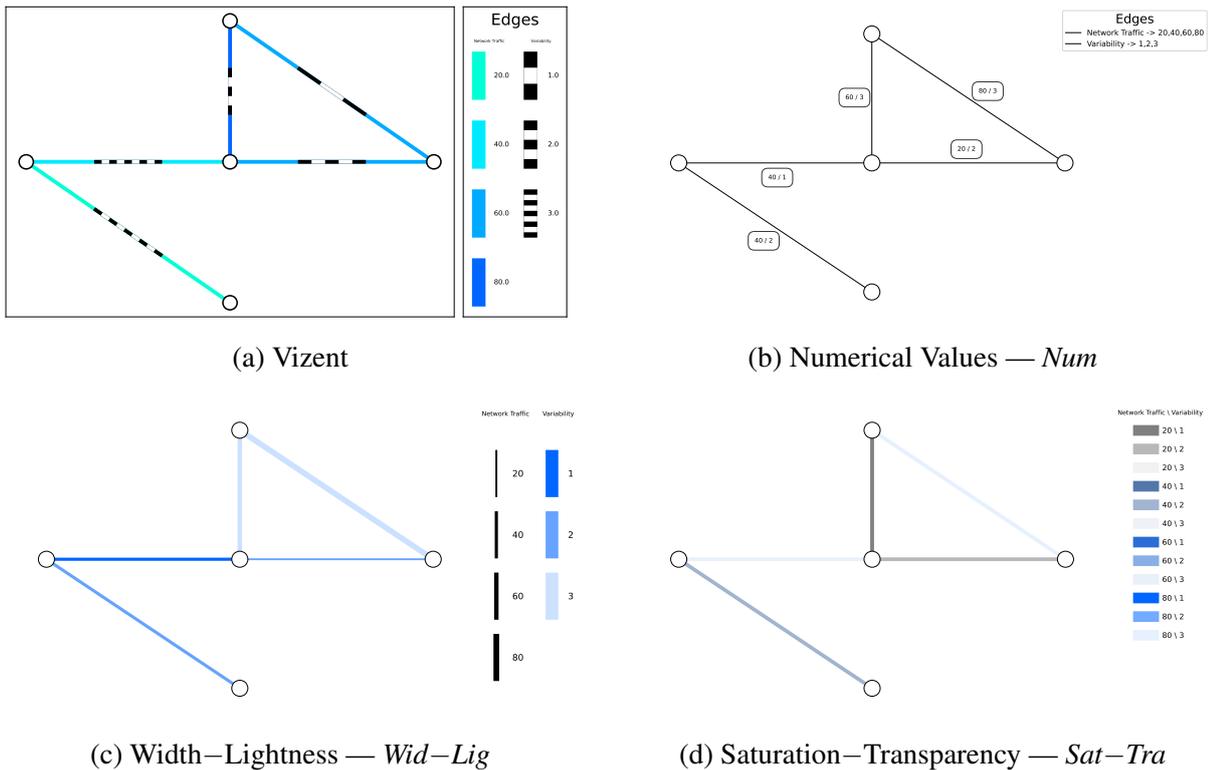


Figure 5.1 Illustrating practice examples of node-link visualisations used in the practice trial. Vizent design (a) and three different visual encodings, (b), (c), and (d), were employed for comparison in our experiment. The graphs provided above have two attributes, each of which is mapped to different visual attributes.

For the control group of visual encodings, see Figure 5.1, width, saturation, and numerical values were encoded as a visual cue for displaying the network traffic attribute, while lightness, transparency, and numerical values were encoded for displaying the variability attribute on the edges of the graph. Each visual encoding was implemented with the shell layout using NetworkX [72] library and saved in PNG format with a resolution of 1280 x 720 pixels, and included a legend showing the meaning of bivariate channels. We employed a white background colour for each stimulus and explicitly designed the graph layout that did not include crossing or overlapping edges for better readability.

We replicated certain visual variables from Guo et al. [71] and Bae et al. [13] to ensure our results could be directly compared with these established findings. We aimed to maintain consistency in both the experimental configuration and the presentation of data. See Figure 5.2 for the default edge visual variables.

Hue	Saturation	Lightness	Granularity	Transparency	Width	Curve
216	100	50	Solid	1 (opaque)	6 pixels	Straight

Figure 5.2 Default Edge Visual Variables of the Experiment.

The colour hue of the Vizent was defined in the HSL (hue-saturation-lightness) colour space. The range of hue was restricted between 170 to 216 (cyan and cyan blue), as in Guo et al. [71]. The variability value was encoded into the striped pattern. The frequency level of the pattern varies with the increasing level of variability in the network data; see Figure 5.3.

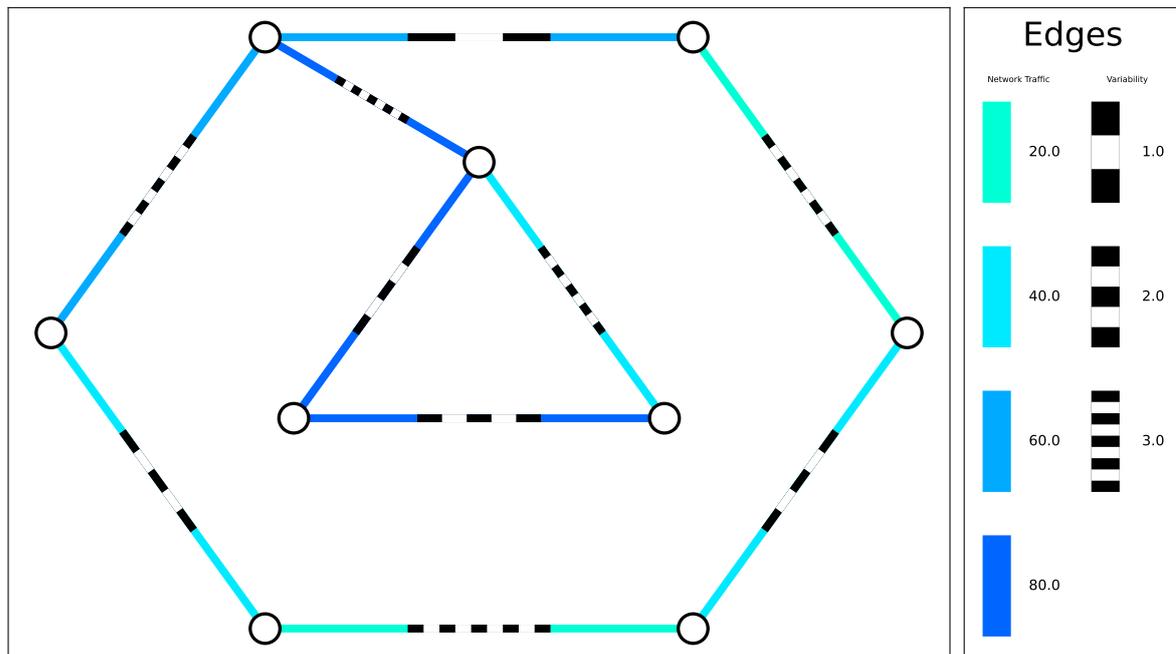


Figure 5.3 A 10-edges graph used as stimuli of Vizent

For *Num*, see Figure 5.4, numerical values were directly used to represent the network traffic value and its variability.

For *Wid-Lig*, the width of edges was varied between 3 pixels to 9 pixels, with thinner edges appearing on lower levels. The lightness number range lies between 0 and 100. In our case, the lightness value ranged from 50 and 90. The hue of the lower-level edges (90) is lighter than that of the higher-level edges (50), see Figure 5.5.

Saturation is defined as a percentage ranging from 0 to 100 (pure colour). For *Sat-Tra*, the saturation level ranged from fully saturated cyan blue (100) to fully desaturated cyan blue (1), which appears grey. The range of transparency values is 0 to 1. While 0 corresponds to a completely transparent fill and 1 to a solid fill. However, in our case, the level of transparency was adjusted between 0.1 (the lowest level) and 1 so that the edge lines are visible, see Figure 5.6.

The experiment provided participants with various visual representations, such as static, undirected, and weighted graphs. Each participant completed 60 trials with different graph layouts: 4 visual encodings x 3 tasks x 5 graph complexity (edge number). The number of edges in the created graphs varied from 5 edges to 25 edges in increments of 5. For each graph size, one

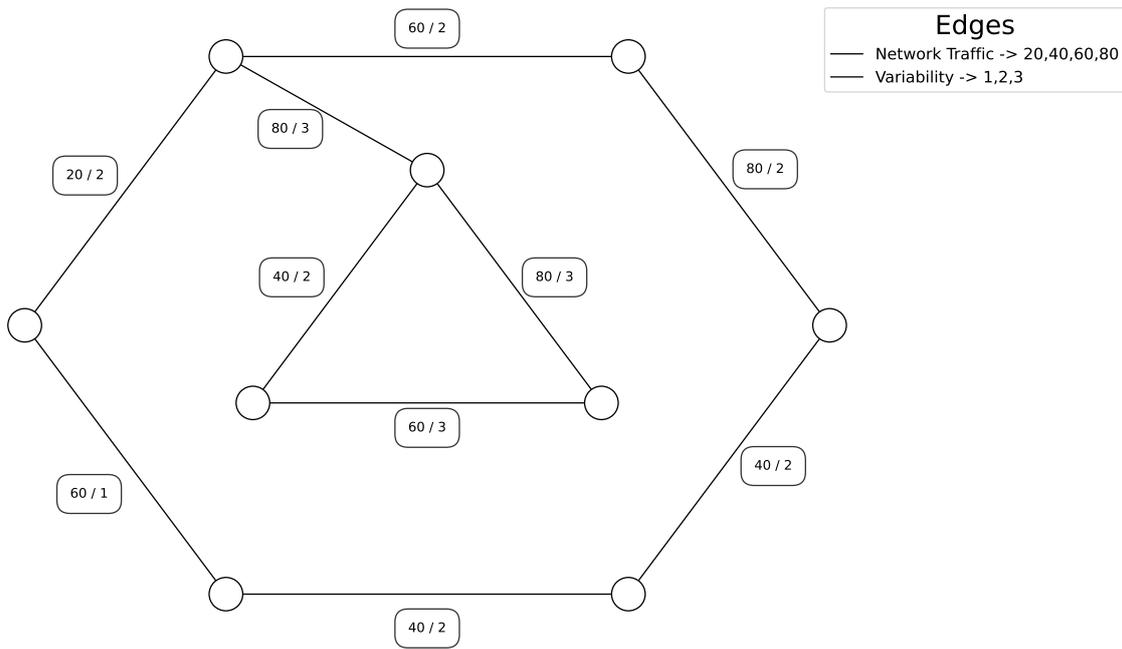


Figure 5.4 A 10-edges graph used as stimuli of *Num*

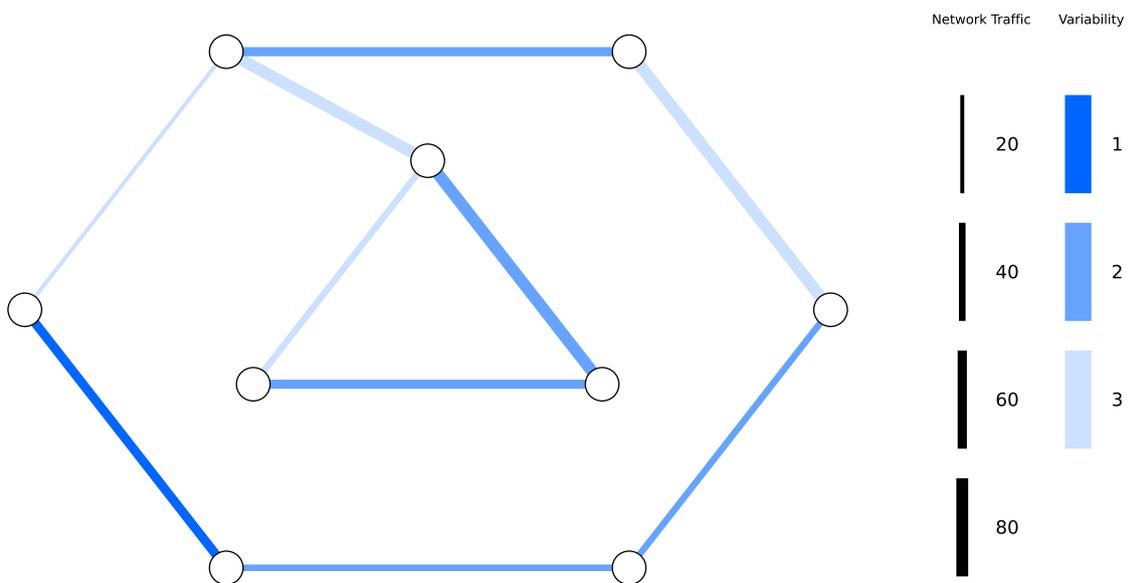


Figure 5.5 A 10-edges graph used as stimuli of *Wid-Lig*

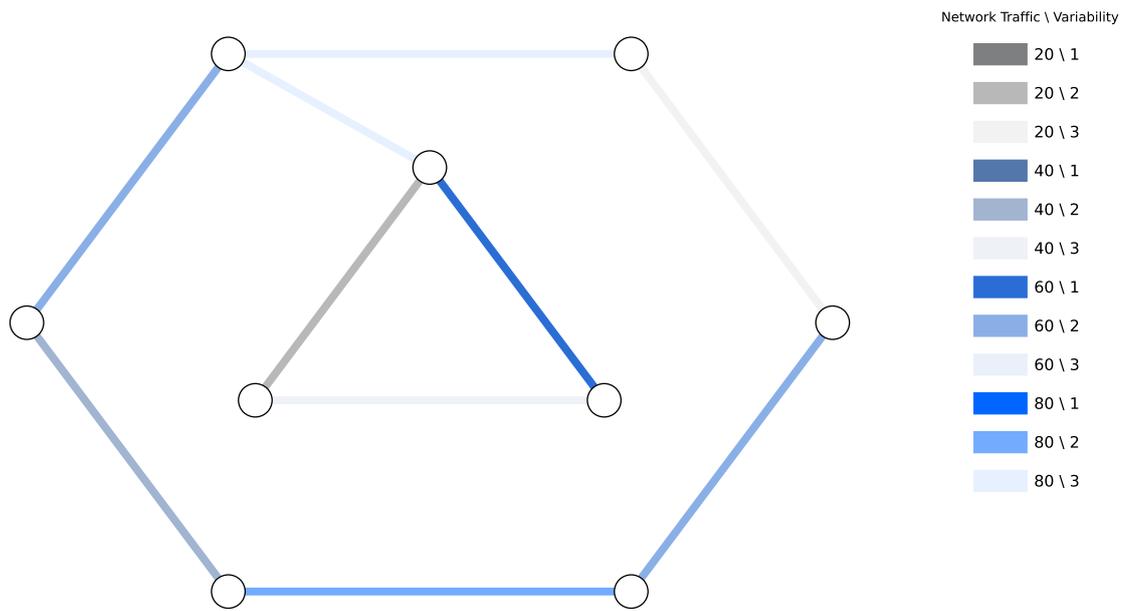


Figure 5.6 A 10-edges graph used as stimuli of *Sat–Tra*

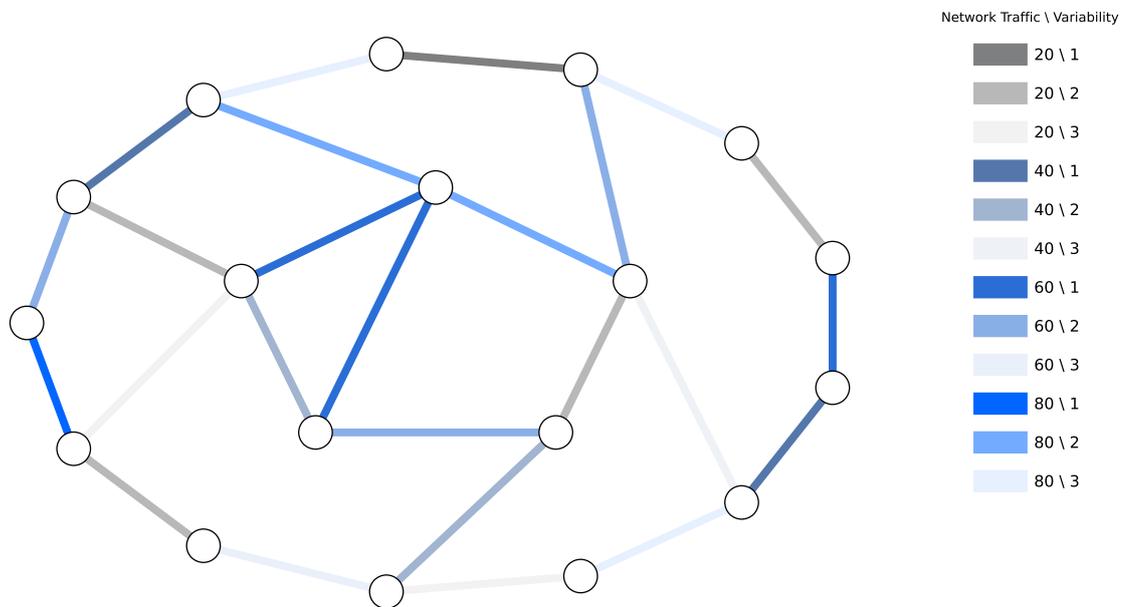


Figure 5.7 The node-link diagram shows an example of a 25-edges graph used as stimuli of *Sat–Tra*.

trial was shown to the participants. A maximum graph size of 25-edges is reasonably depicted on a standard-size screen without zooming or panning, see Figure 5.7. Examples of 15-edges, 20-edges, and 25-edges graphs used as stimuli of all visual encodings in the experiment can be seen in Appendix A.

5.2.3 Tasks

The experiment consisted of three tasks. All participants were required to complete three different tasks for four visual encodings. The order of tasks was randomised to reduce the learning effect, so each subject performed a different shuffled order of tasks. Participants were not informed about how Vizent design works or which strategy we proposed to avoid inadvertently influencing them. As a result, we are purposefully vague: we just asked them to “find the targeted edge in the context of given visual encoding”.

Lee et al. [113] provide a comprehensive summary of graph exploration tasks and differentiate between topology-based tasks, attribute-based tasks, and browsing and overview tasks. Attribute-based tasks include the identification of nodes or links with certain labels. The browsing tasks encompass several topological operations, including examining adjacency relationships between nodes, determining the shortest path between nodes, and the identification of clusters within the graph. On the other hand, the overview task focuses on examining the network’s general attributes, such as estimating its overall size. The topology-based tasks hold greater significance in navigating a graph’s structure. However, in our experiment, we do not place emphasis on the graph’s topology.

For this experiment, attribute-based tasks were chosen and inspired by our detailed inquiry into the issues surrounding server interconnections and their functionality and use cases. Real Facebook traffic patterns [5] were explored to understand important questions about the network data, features, and relationships between entities, including the size of the flows and bandwidth variations through time. These high-level tasks typically involve performing several low-level tasks [113].

Table 5.1 List of experimental questions used in this study.

<i>Attribute-based tasks</i>	<i>Task Type</i>	<i>Target</i>	<i>Visual Variable</i>
<i>Task 1) Find the edge with the lowest variability in the graph.</i>	<i>A single visual attribute</i>	<i>Single</i>	<i>Variability</i>
<i>Task 2) Find the edge with the highest network traffic in the graph.</i>	<i>A single visual attribute</i>	<i>Single</i>	<i>Network Traffic</i>
<i>Task 3) Find the edge with the highest network traffic and lowest variability.</i>	<i>Two visual attributes</i>	<i>Dual</i>	<i>Network traffic and Variability</i>

Our tasks represent a variety of realistic network tasks while covering the basic perceptual operations of distinguishing high-level “interesting” edge patterns between relational nodes and identifying a targeted edge having an extremum value. All formed tasks were carefully chosen to satisfy both Lee et al.’s graph task taxonomy [113] and Amar et al.’s taxonomy of low-level visualisation tasks [4]. All experimental questions were described in Table 5.1.

Two separate target types were created for the tasks: *Single* and *Dual*. For *Single* target, subjects are asked to identify the one visual variable on edge: network traffic or the variation in that traffic. However, for *Dual* target (*Task 3*), they ought to search the target features on the edges by considering the two variables of the graph edge: network traffic and its variability.

After generating the datasets, each one was thoroughly examined to verify that it included only one specific data needed for its designated task. By taking this measure, it was guaranteed that each task would have only one correct answer, thereby preserving the integrity of the task-specific data utilised in the experimental design and guaranteeing the correctness of the experiment.

5.2.4 Hypotheses

In this experiment, we devise and analyse the following null hypotheses:

H1 There is no statistically significant difference in response time (efficiency) and accuracy (effectiveness) between Vizent and the control group of visual encodings when completing *Task 1*.

H2 There is no statistically significant difference in response time and accuracy between Vizent and the control group of visual encodings when completing *Task 2*.

H3 There is no statistically significant difference in response time and accuracy between Vizent and the control group of visual encodings when completing *Task 3*.

The null hypotheses (**H1**, **H2**, **H3**) were defined, and they do not specify whether Vizent performs better or worse than the control group of visual encodings across the tasks. The principal rationale behind this approach is to ensure objectivity in our experimental analysis. By establishing our hypotheses in this manner, we do not imply that Vizent is superior or inferior to alternative encoding methods. On the contrary, our objective is to determine whether Vizent performs equivalently in terms of response time (efficiency) and accuracy (effectiveness) to existing visual encoding. Moreover, employing this null hypothesis approach aids in mitigating bias in interpreting the results since it places the burden of proof on demonstrating a difference rather than assuming it initially. A statistically significant difference would be indicated if evidence were to support the rejection of these null hypotheses, thus implying that Vizent either enhances or diminishes performance in comparison control groups.

5.2.5 Experimental Procedure

Python 3.9 and PschoPy version 2021 [142] were used to design the experiment. We conducted a within-subject experiment to eliminate the influence of personal interests and abilities through the Prolific [140] online experiment tool. For the study, 50 participants with normal or corrected vision were recruited. This experiment did not consider the demographic data (gender, age,

educational background). All the participants were native English speakers and paid 2.50 pounds to complete the experiment (intended to take less than 20 min). We did not impose a time limit on the tasks, but we advised participants to complete the tasks as quickly and accurately as possible without assigning any priority between the two performance metrics.

Participants were required to use a laptop or desktop computer with a resolution of at least 1366 x 768 pixels available screen space to take part in the experiment. All participants agreed to a Newcastle University-approved consent form and, after that, read the learning task instructions carefully.

A brief short trial session was implemented for each of the four visual encodings in the experiment. Thus, participants performed only two practice trials for each visual encoding before conducting a full experiment. During trial sessions, the evaluation software did not gather the data; rather, it displayed the correct answer upon completion of each task. Immediate results were displayed, showing whether the result was correct or not after each trial.

5.2.6 Edge Performance Experiment Results

Participants' accuracy and response time were recorded during each trial, enabling the evaluation of two different performance measures. Two dependent variables, the response time (in seconds) and accuracy, were used to evaluate the effectiveness of each visual encoding. There was only one correct response for each trial, and the task accuracy was measured as the proportion of the number of correct replies to the number of stimuli within each task.

Before analysis, we applied a data quality check separately to the response time and accuracy of the four visual encodings: *Vizent*, *Num*, *Wid-Lig*, and *Sat-Tra*. The z-score was calculated specifically for each participant's average accuracy across all visual encodings and tasks, and these scores were compared to a threshold of -3 to +3 standard deviations from the mean. We identified and removed one participant whose average accuracy was close to chance, indicating potential misunderstanding of the tasks or random clicking. The response time check revealed that some users took unusually long to finish tasks. After reviewing the results of each task for all visual encodings, we concluded that they spent this time locating the specified edge.

Following the quality checks, we analysed the data of participants' response time and accuracy separately. A Shapiro-Wilk method with a significance level of $\alpha = 0.05$ was run on each visual encoding to determine whether or not the data was normally distributed. In addition, we plotted the distributions of the data to examine normality. The Shapiro-Wilk test indicated violations of the assumptions of normality for response time and accuracy.

We analyse the results by treating each task as a separate experiment. For the non-normally distributed accuracy and response time, we applied a Friedman non-parametric analysis with a significance level of $\alpha = 0.05$ to determine if there was a significant difference between the four visual encodings within each *Task*. This was followed by a post hoc analysis using a Wilcoxon signed-rank test with Bonferroni correction for pairwise comparison.

The Bonferroni correction is an adjustment made to the p-value in order to reduce the chances of obtaining false-positive outcomes (protects from Type I errors) in the context of multiple pairwise tests conducted on a singular dataset [134]. The Bonferroni correction involves a

reduction of the individual significance level (p-value) to “ p/n ” (where n represents the number of comparisons being conducted). Post hoc pairwise comparisons with the Bonferroni correction (at an adjusted $p = 0.0125$) were performed to determine significant differences between pairs of visual encodings regarding response time and accuracy.

The results of the experiment are presented according to the three tasks evaluated. Table 5.2 below represents the post hoc (z -, p - values) pairwise analyses for each task with significant differences in accuracy and response time highlighted in bold using $p = 0.0125$ following Bonferroni correction. If statistically significant differences were found between pairs of visual encodings, we additionally calculated r -value (effect size) ($r = Z/\sqrt{N_{pairs}}$), recommended in [157], as shown in Table 5.2. The effect size (r) is categorized into three groups using Cohen [41] criteria: $0.1 < r < 0.3$ being a small effect, $0.30 < r < 0.50$ being a medium effect, and finally $r \geq 0.50$ being a large effect.

Table 5.2 Post hoc analysis results with Wilcoxon signed rank tests for each *Task*. Significant differences between pairs of visual encodings are highlighted in bold using $p < 0.0125$.

		<i>Vizent vs</i> <i>Num</i>	<i>Vizent vs</i> <i>Wid–Lig</i>	<i>Vizent vs</i> <i>Sat–Tra</i>	<i>Num vs</i> <i>Wid–Lig</i>	<i>Num vs</i> <i>Sat–Tra</i>	<i>Wid–Lig vs</i> <i>Sat–Tra</i>	
<i>Task 1</i>	Accuracy	z	-0.683	0.517	2.134	1.239	2.646	1.935
		p	0.494	0.605	0.033	0.215	0.008	0.053
		r	-	-	-	-	0.27	-
	Response Time	z	-3.049	4.461	-0.761	5.237	1.219	-4.034
		p	0.002	0.000	0.447	0.000	0.223	0.000
		r	0.31	0.45	-	0.53	-	0.41
<i>Task 2</i>	Accuracy	z	-0.121	5.418	5.853	5.53	5.904	3.41
		p	0.904	0.000	0.000	0.000	0.001	0.001
		r	-	0.55	0.59	0.55	0.60	0.34
	Response Time	z	-4.571	-1.915	-3.059	3.288	1.616	-2.333
		p	0.000	0.056	0.002	0.001	0.106	0.020*
		r	0.46	-	0.31	0.33	-	-
<i>Task 3</i>	Accuracy	z	-2.483	2.473	0.782	4.553	3.432	-1.43
		p	0.013*	0.013*	0.043	0.000	0.001	0.153
		r	-	-	-	0.46	0.35	-
	Response Time	z	-0.642	3.974	3.815	3.994	4.541	-1.149
		p	0.521	0.000	0.000	0.000	0.000	0.251
		r	-	0.40	0.38	0.40	0.46	-

* indicates a p-value close to the significance threshold (0.0125), which could be influenced by the strictness of the Bonferroni correction [10].

Figure 5.8 illustrates time and accuracy by all visual encodings, grouped by *Task*; boxplots represent response time with the lower, median, and upper quartiles, while accuracy is represented by bars (mean) and error bars (95 confidence intervals). The black lines that connect pairs of visual encodings indicate statistically significant differences across visual encodings.

T1 (Variability) — For the first experimental question: *Find the edge with the lowest variability in the displayed graph.*

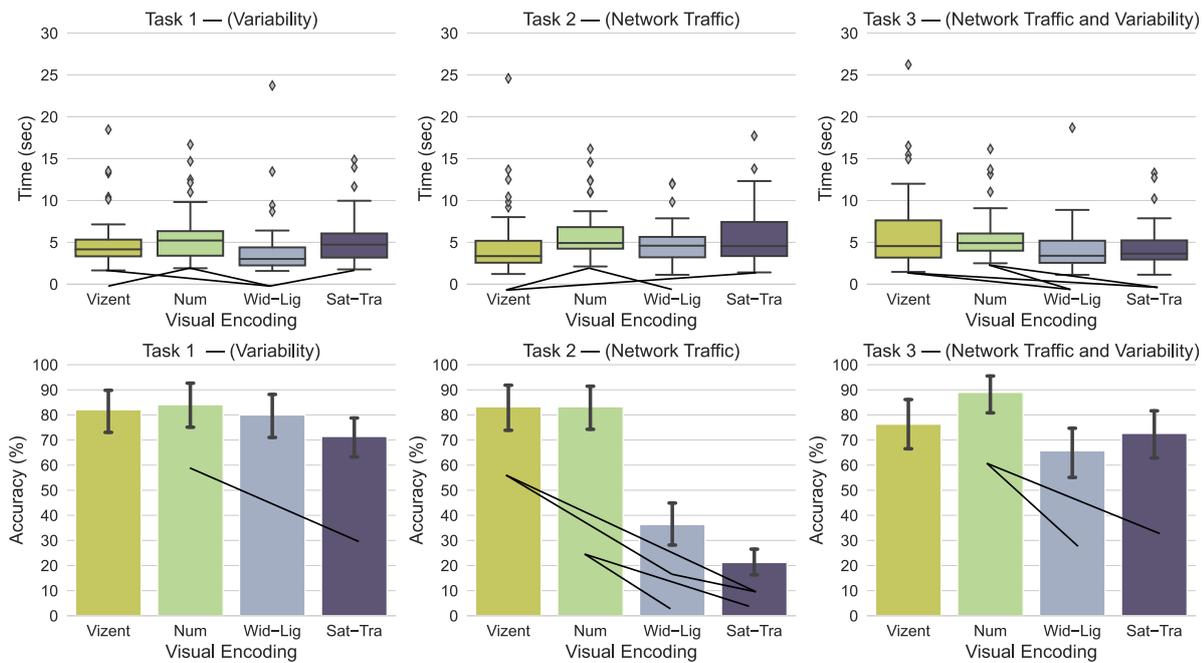


Figure 5.8 Response times in seconds (upper side, as box plots) and accuracy (bottom side, as bars representing mean values and error bars representing 95% confidence intervals) classified by Vizent, *Num*, *Wid-Lig*, *Sat-Tra*, respectively grouped by *Task*. The black lines highlight significance between visual encodings.

The Friedman test revealed significant differences in response time among the four visual encodings ($X^2 = 47.180$, $df = 3$, $p < 0.001$). Following a post hoc analysis, it was revealed that the *Wid-Lig* was significantly faster than Vizent, *Num*, and *Sat-Tra* ($p < 0.001$). Also, a significant difference was found between Vizent and *Num* in favour of Vizent ($p < 0.01$).

Regarding accuracy, we also find a significant difference among the four visual encodings ($X^2 = 19.813$, $df = 3$, $p < 0.001$). The post hoc analysis revealed that *Num* is significantly more accurate than *Sat-Tra* ($z = 2.646$, $p < 0.001$) with a small effect size ($r = 0.27$). However, there was no significant difference found between the following pairings: Vizent/*Wid-Lig*; Vizent/*Num*; Vizent/*Sat-Tra*; *Num*/*Wid-Lig*; *Wid-Lig*/*Sat-Tra*.

Key findings: The resulting data suggest that our first hypothesis **H1** holds in terms of accuracy; however, it does not hold in terms of response time as *Wid-Lig* produced quicker response time than Vizent, and the subjects answered significantly faster with Vizent than in *Num* for *Task 1*.

Task 1 discussion: In general, the results of *Task 1* indicate that participants responded more quickly to *Wid-Lig* than the other visualisation approaches. Our findings are not surprising for Vizent, considering we presented a novel edge design to participants for the first time, requiring additional training. As a result, it is reasonable that they spent more time locating the targeted edge with Vizent than with *Wid-Lig*. Additionally, Vizent produced a quicker response time than both *Num* and *Sat-Tra*. The medians indicated that while the shortest response time was spent with the *Wid-Lig* ($\tilde{x} = 3.01$ sec.), the longest time was spent with *Num* ($\tilde{x} = 5.22$ sec.). A shorter time was spent with Vizent ($\tilde{x} = 4.16$ sec.) than with *Sat-Tra* ($\tilde{x} = 4.72$ sec.), see Table 5.3.

Table 5.3 The mean and median response times in seconds are grouped by visual encoding and *Task*. The lowest mean response time of the visual encodings is highlighted in bold for each *Task*. The terms *Single* and *Dual* target visual search are also displayed.

			<i>Visual Encodings</i>			
			<i>Vizent</i>	<i>Num</i>	<i>Wid–Lig</i>	<i>Sat–Tra</i>
<i>Task 1</i>	<i>Single</i>	Mean	5.03	5.76	4.07	5.25
		Median	4.16	5.22	3.01	4.72
<i>Task 2</i>	<i>Single</i>	Mean	4.63	6	4.79	5.74
		Median	3.35	4.91	4.59	4.55
<i>Task 3</i>	<i>Dual</i>	Mean	6.31	5.64	4.35	4.41
		Median	4.55	4.91	3.39	3.64

In terms of accuracy, a post hoc analysis revealed a significant difference between *Num* and *Sat–Tra* in favour of *Num* ($p = 0.008$), indicating that participants' accuracy was significantly lowered with *Sat–Tra* when answering *Task 1*. In particular, the difference between the accuracy of *Num* and *Sat–Tra* is about 15%. However, there was no significant difference found between the remaining pairs.

T2 (Network Traffic) — For the second experimental question: *Find the edge with the highest network traffic in the displayed graph.*

The Friedman test revealed significant differences among the four visual encodings in terms of response time ($X^2 = 39.588$, $df = 3$, $p < 0.001$) and accuracy ($X^2 = 91.349$, $df = 3$, $p < 0.001$). The post hoc analysis revealed that *Vizent* is significantly faster than *Num* and *Sat–Tra* regarding the response time ($p < 0.001$) with medium effect size, shown in Table 5.2. Additionally, we found that *Wid–Lig* is significantly faster than *Num*.

Regarding accuracy, the post hoc analysis revealed that *Vizent* is significantly more accurate than *Wid–Lig* and *Sat–Tra* ($p < 0.001$) with a large effect size ($r > 0.5$). Similarly, *Num* is significantly more accurate than *Wid–Lig* and *Sat–Tra* visualisations ($p < 0.001$) with a large effect size ($r > 0.5$). We also found that *Wid–Lig* is significantly more accurate than *Sat–Tra*.

Key findings: *Vizent* produced both more accurate and quicker performance than both *Wid–Lig* and *Sat–Tra*, and significantly faster response time than *Num*. Therefore, these results falsified the second hypothesis **H2** on both accounts: response time and accuracy.

Task 2 discussion: In general, we found a discrepancy in the findings between *Num* and *Wid–Lig* (*Wid–Lig* is faster than *Num* yet *Num* is significantly more accurate than *Wid–Lig*). For this reason, a correlation analysis was conducted response between time and accuracy for *Wid–Lig*, and the analysis result was positive but not statistically significant. As a result, we might conclude that faster responses do not necessarily result in lower accuracy for *Wid–Lig*.

Vizent showed better task performance than other visual encodings in *Task 2*. Participants answered more accurately and spent less time when working with *Vizent* than both *Wid–Lig* and *Sat–Tra*, and produced quicker response time than *Num*. Regarding accuracy, *Vizent* and *Num* performed similarly well in *Task 2* with the same accuracy of about 83%; however, the medians

revealed that *Num* ($\bar{x} = 4.91$ sec.) took the longest response time, while Vizent produced the quickest performance ($\bar{x} = 3.53$ sec.).

Wid–Lig and *Sat–Tra* underperformed on this task, with *Sat–Tra* performing the lowest. Lightness and transparency may interfere with accurate visual perception, as demonstrated by the intriguing finding. Particularly, we may deduce that transparency strongly interacts with saturation, reducing the ability of the respondents to distinguish the fully saturated edge, and should not be used in conjunction with it.

T3 (Network Traffic and Variability) — For the third experimental question: *Find the edge with the highest network traffic and lowest variability in the displayed graph.*

The Friedman test revealed significant differences in response time among the four visual encodings ($X^2 = 42.233$, $df = 3$, $p < 0.001$). The post hoc analysis revealed that *Wid–Lig* is significantly faster than Vizent and *Num*. Similarly, *Sat–Tra* is significantly faster than Vizent and *Num*.

We also find significant differences in accuracy ($X^2 = 32.341$, $df = 3$, $p < 0.001$). Four pairwise comparisons revealed that *Num* is significantly more accurate than *Wid–Lig* ($p < 0.001$) and *Sat+Tra* ($p < 0.001$) with medium effect size. However, no significant differences were found between the following pairings: Vizent/*Wid–Lig*; Vizent/*Num*; Vizent/*Sat–Tra*; *Wid–Lig*/*Sat–Tra*. On the other hand, the results of Vizent versus *Num* ($p = 0.013$), and Vizent versus *Wid–Lig* ($p = 0.013$) are close to the significance level, which could be indicative of a type-2 error resulting from Bonferroni adjustment being too strict [10].

Key findings: The analysis results for confirming the third hypothesis **H3** were similar to those reported in *Task 1*. We reject the **H3** for response time as the response time differed significantly with the exception of Vizent versus *Num*. Again, no significant difference in accuracy was found between Vizent and the other visual encodings; therefore, we fail to reject the **H3** concerning accuracy.

Table 5.4 The mean and median accuracy (%) are grouped by visual encoding and *Task*. The highest mean accuracy percentage of the visual encodings is highlighted in bold for each *Task*. The terms *Single* and *Dual* target visual search are also displayed.

			<i>Visual Encodings</i>			
			<i>Vizent</i>	<i>Num</i>	<i>Wid–Lig</i>	<i>Sat–Tra</i>
<i>Task 1</i>	<i>Single</i>	Mean	82.04	84.08	80	71.42
		Median	100	100	100	80
<i>Task 2</i>	<i>Single</i>	Mean	83.26	83.26	36.32	21.22
		Median	100	100	40	20
<i>Task 3</i>	<i>Dual</i>	Mean	76.32	88.97	65.71	72.65
		Median	100	100	80	80

Task 3 discussion: Regarding the response time, the post hoc analysis revealed that both *Wid–Lig* and *Sat–Tra* are significantly faster than *Num* and Vizent. However, due to the discrepancy in the findings for *Num* and both the *Wid–Lig* and *Sat–Tra* (*Wid–Lig* and *Sat–Tra* are

faster than *Num*, yet *Num* produces significantly more accurate performance than *Wid–Lig* and *Sat–Tra*), we conducted a correlation analysis between time and accuracy for both the *Wid–Lig* and *Sat–Tra*. The outcomes were negative and statistically significant for both visualisation methods. This revealed that faster responses did not result in lower accuracy.

In terms of accuracy, *Num* had a better accuracy of approximately 90%. When working with *Vizent*, the participants achieved a second higher level of accuracy of around 77%, compared to *Wid–Lig* (around 66%) and *Sat–Tra* (around 73%), see Table 5.4 above. It is worth noting that no significance was found between *Vizent* and *Num* following the Bonferroni correction.

5.2.7 Graph Size

As varying graph size is another interest of this study, we formulated the following secondary research question “Does the varying number of graph edges significantly influence the performance of each visual encoding?”. The participants were provided with various graphs ranging in complexity from 5 to 25 edges in increments of 5 to observe how increasing the size of graphs affects each visual encoding regarding performance measures. In this case, the Bonferroni-corrected p – value was found by dividing the p-value by the number of the graph sizes ($0.05/5 = 0.01$), and then we applied ($p = 0.01$) to our results as the significance level.

We summarised the effects of graph complexity in edge size within the provided task in terms of response time and accuracy. It should be noted that the graph size at its largest is not huge and that this corresponds to other works in [71, 13] rather than an upper limit of the possible.

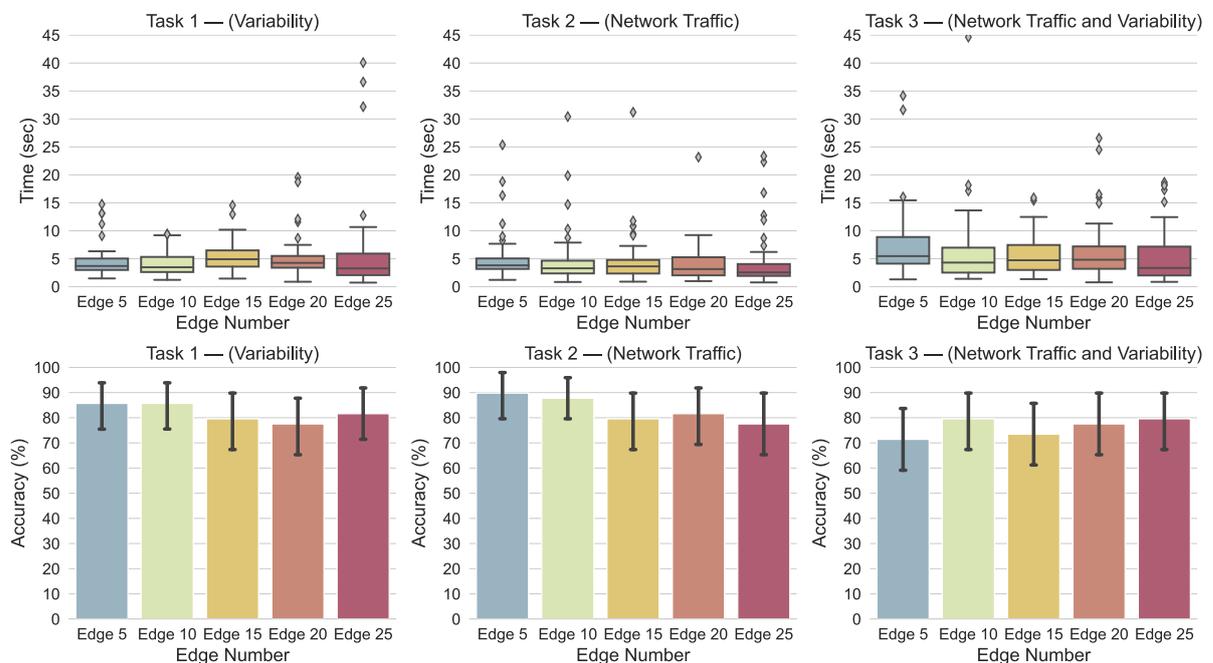


Figure 5.9 Response times in seconds (upper side, as box plots) and accuracy (bottom side, as bars representing means and error bars representing 95% confidence intervals) classified by edge number, respectively grouped by *Task* within *Vizent*.

Within *Vizent*: The obtained results are illustrated in Figure 5.9 for each *Task* separately. For *Task* 1, we found significant differences in terms of response time ($X^2 = 21.94$, $df = 4$,

$p < 0.001$), however, no significant difference was found in terms of accuracy ($p = 0.43$) between the five various levels of graph edges. Post hoc analysis revealed that participants only responded significantly faster to the 10-edges than to the 15-edges.

For *Task 2*, we found significant differences in terms of response time ($X^2 = 29.56, df = 4, p < 0.001$), and accuracy ($X^2 = 9.57, df = 4, p < 0.05$) between the five various levels of graph edges. Post hoc analysis revealed that participants responded significantly faster to the 10-edges, 20-edges, and 25-edges than to the 5-edges. Also, they responded significantly faster to the 25-edges than to the 15-edges. Regarding accuracy, we observed no statistically significant difference following the Bonferroni correction.

For *Task 3*, we found significant differences in terms of response time ($X^2 = 19.88, df = 4, p < 0.01$), however, no significant difference was found in terms of accuracy ($p = 0.33$) between the five various levels of graph edges. Post hoc analysis revealed that participants responded significantly faster to the 10-edges and 25-edges than to the 5-edges.

Key findings: For all tasks, we found no notable significant evidence that increasing the graph size has a discernible influence on response time and accuracy within Vizent.

Within Num: The obtained results are illustrated in Figure 5.10 for each *Task* separately. For *Task 1*, we found significant differences in terms of response time ($X^2 = 24.01, df = 4, p < 0.001$), however, no significant difference was found in terms of accuracy ($p = 0.83$) between the five various levels of graph edges. Post hoc analysis revealed that participants spent significantly more time on 20- and 25-edges than on up to 15-edges when looking for the target.

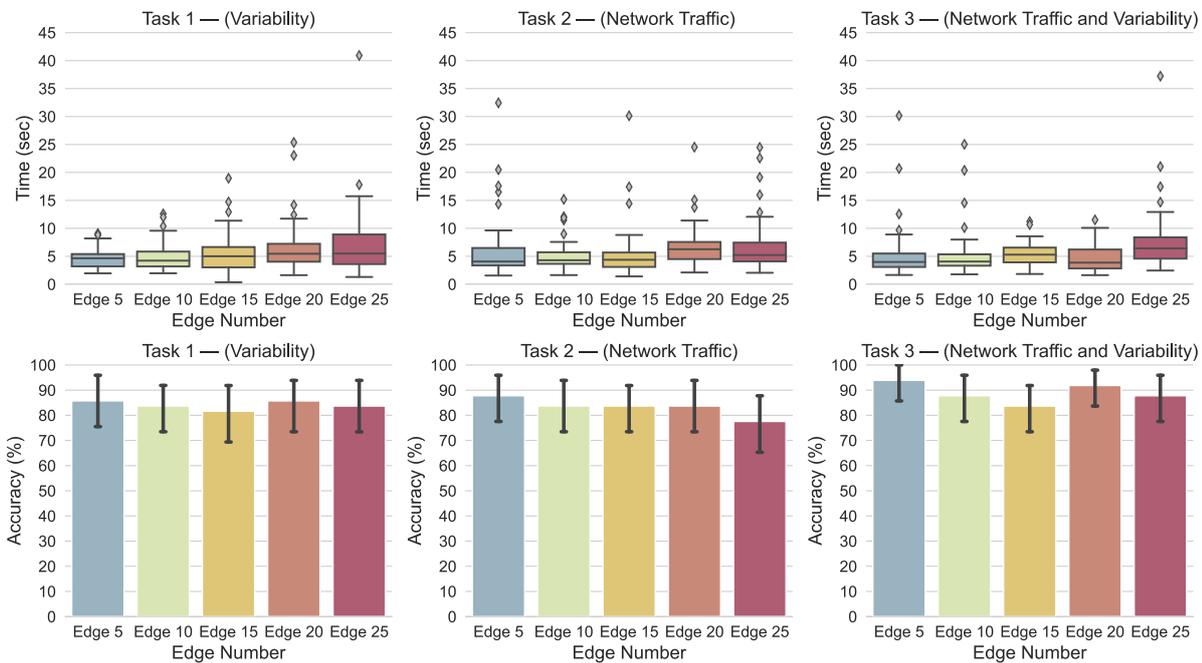


Figure 5.10 Response times in seconds (upper side, as box plots) and accuracy (bottom side, as bars representing means and error bars representing 95% confidence intervals) classified by edge number, respectively grouped by *Task* within *Num*.

For *Task 2*, we found significant differences in terms of response time ($X^2 = 28.90, df = 4, p < 0.001$), however, no significant difference was found in terms of accuracy ($p = 0.33$) between the five various levels of graph edges. Post hoc analysis revealed that participants spent

significantly more time on 20- and 25-edges than on up to 15-edges, except for a slight difference between 5- and 25-edges.

For *Task 3*, we found significant differences in terms of response time ($X^2 = 36.27$, $df = 4$, $p < 0.01$), however, no significant difference was found in terms of accuracy ($p = 0.07$) between the five various levels of graph edges. Post hoc analysis revealed that participants only spent significantly more time on 25-edges than on up to 20-edges, but there were no statistically significant differences between the remaining edges.

Key findings: Our results are intriguing in that participants' response time significantly increases when the graph size exceeds 15-edges or 20-edges depending on the task type. For *Task 1* and *Task 2*, we found that increasing graph size beyond 15-edges significantly affected response time. However, for *Task 3*, participants' response time significantly increases when the graph size reaches 25. As a result, it is possible to infer that locating the desired edge in *Num* will require participants to spend more time hunting around the entire graph design as edge size increases beyond 15.

Within *Wid–Lig*: The obtained results are illustrated in Figure 5.11 for each *Task* separately.

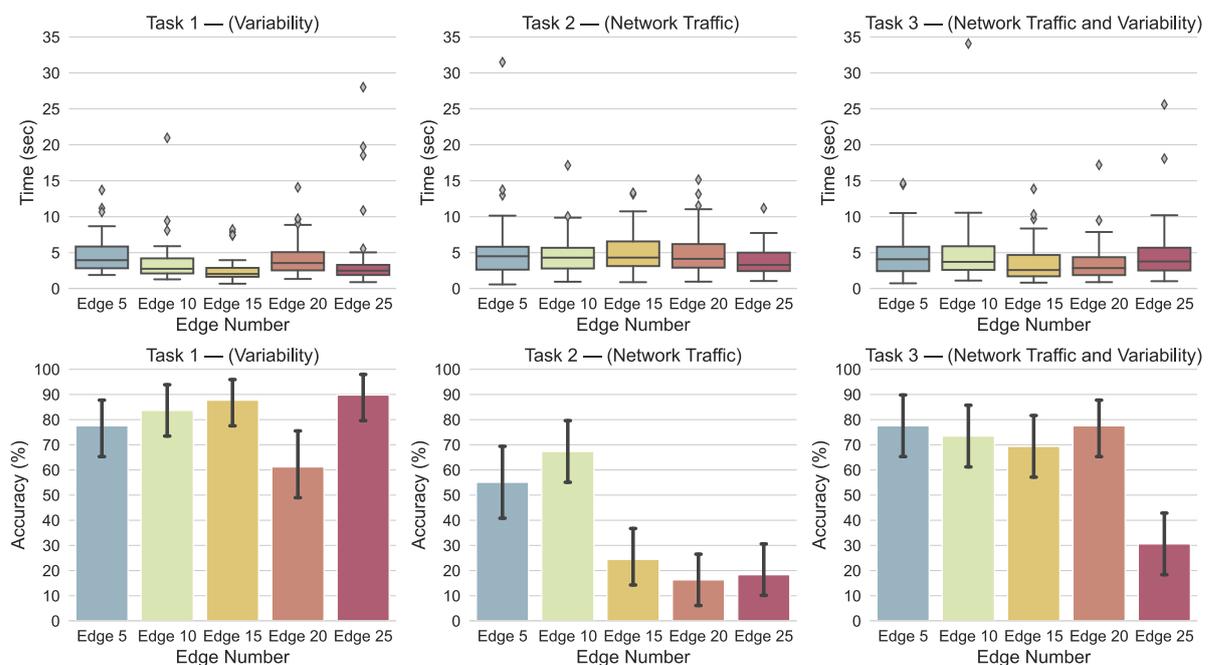


Figure 5.11 Response times in seconds (upper side, as box plots) and accuracy (bottom side, as bars representing means and error bars representing 95% confidence intervals) classified by edge number, respectively grouped by *Task* within *Wid–Lig*.

For *Task 1*, we found significant differences in terms of response time ($X^2 = 76.52$, $df = 4$, $p < 0.001$), and accuracy ($X^2 = 32.51$, $df = 4$, $p < 0.01$) between the five various levels of graph edges. Participants significantly spent more time on the 5- and 20-edges than on the remaining edges during this task. Also, their accuracy dropped significantly to 60% when the edge number was 20 compared to the remaining edges. However, according to the post hoc analysis, no significant difference in accuracy was found between 5-edge and 20-edge. The decrease in accuracy when the graph size was 20 could be due to the primary encoding value,

where network traffic was lowest, so the target edge was marked as a thin line, limiting the discriminability of the lightness of the visual channel for the target edge.

For *Task 2*, we found no significant differences in terms of response time ($p = 0.19$). However, significant differences in accuracy were found ($X^2 = 61.50, df = 4, p < 0.001$). We discovered that beyond 10-edges, participants' accuracy fell significantly to around 20%. They responded with an average accuracy of 60% within 5- and 10-edges, which is still poor performance. After a thorough analysis of the results, it is worth noting that a significant fall in accuracy could imply that second visual encoding (varying lightness level) might influence the saliency of the targeted edge in *Task 2*.

For *Task 3*, we found significant differences in terms of response time ($X^2 = 12.53, df = 4, p < 0.05$), and accuracy ($X^2 = 62.42, df = 4, p < 0.01$). The participants' accuracy decreased significantly to around 30% within 25-edges; however, no significant differences were found across the remaining edges.

Key findings: For all tasks, no clear evidence of a diminishing or increasing tendency in response time was found for different graph sizes. It is not possible to generalize that the complexity of the graph influences the participant's accuracy for all tasks; however, for *Task 3*, participants' accuracy decreases dramatically beyond 20-edges. Regarding accuracy, we can conclude that width and lightness do not work well together as lightness level, to some degree, visually interfere with the width of the edge for *Task 2*.

Within Sat–Tra: The obtained results are illustrated in Figure 5.12 for each *Task* separately.

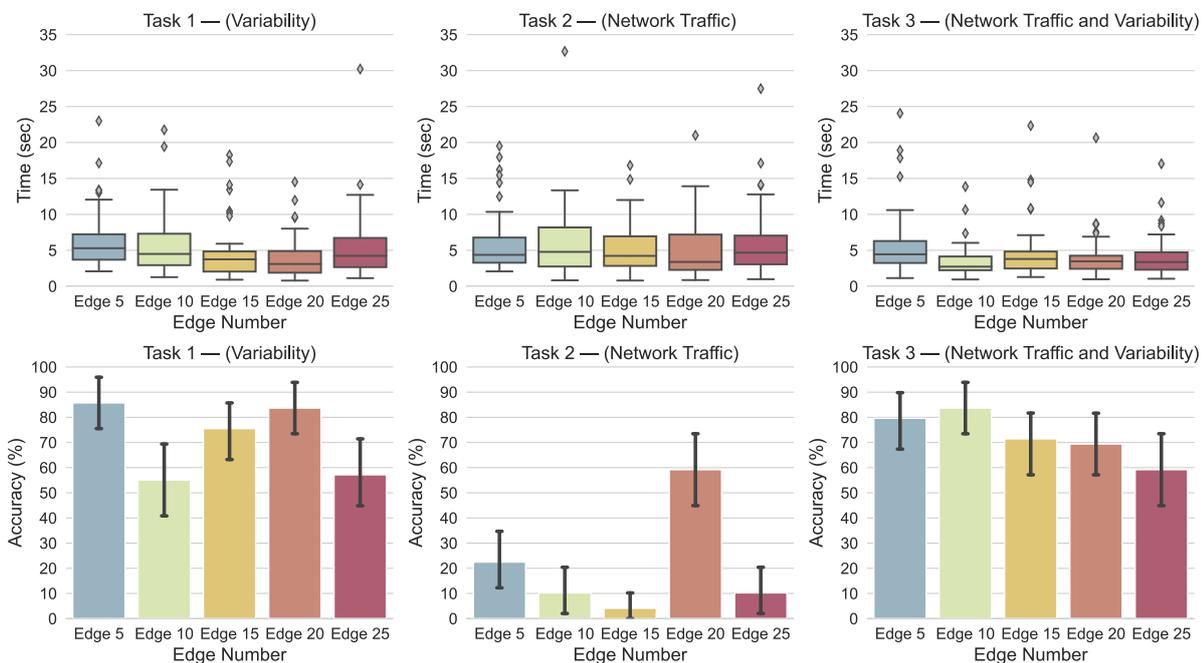


Figure 5.12 Response times in seconds (upper side, as box plots) and accuracy (bottom side, as bars representing means and error bars representing 95% confidence intervals) classified by edge number, respectively grouped by *Task* within *Sat–Tra*.

For *Task 1*, we found significant differences in terms of response time ($X^2 = 27.83, df = 4, p < 0.001$), and accuracy ($X^2 = 26.57, df = 4, p < 0.001$) between the five various levels of graph edges. By contrast, these statistically significant results did not conclusively establish that

increasing the size of the edge affects the outcomes. Nonetheless, the difference is noteworthy, so we decided to perform additional investigations. As shown in Figure 5.12, when the edge size was 10 or 25, the participants' accuracy declined to 55%. Additionally, participants spent significantly less time on the 15- and 20-edges than on the remaining edges. We discovered that the value of network traffic was a significant factor in achieving higher accuracy and quicker response time. In other words, when the network value of the target edges was at its lowest or highest (fully desaturated or fully saturated), the participants located the edge with higher accuracy and in less time; otherwise, their accuracy dropped, and they required more time.

For *Task 2*, we found significant differences in terms of response time ($X^2 = 12.53$, $df = 4$, $p < 0.001$), and accuracy ($X^2 = 57.91$, $df = 4$, $p < 0.001$) between the five various levels of graph edges. The post hoc analysis revealed that participants responded significantly faster to the 20-edges than to the 25-edges. Also, they provided significantly higher responses to the 20-edges with an accuracy of 60% than the remaining edges. This significant difference indicates that variations in the transparency level had a statistically significant negative effect on accuracy when finding the most network value. More interestingly, they provided a more accurate response when the variability value of the targeted edge was the lowest (the highest transparency); otherwise, they reached only about 20% accuracy.

For *Task 3*, we found significant differences in terms of response time ($X^2 = 28.53$, $df = 4$, $p < 0.001$), and accuracy ($X^2 = 19.37$, $df = 4$, $p < 0.01$) between the graph sizes. The post hoc analysis revealed that participants spent significantly more time on the 5-edges than on the 10- and 15-edges. However, no significant difference in response time was found amongst the remaining pairs following a Bonferroni correction.

Key findings: The increased graph complexity in size does not consistently appear to influence the response time and accuracy significantly; however, we noticed that this visual encoding significantly affected the outcomes for *Task 1* and *Task 2*.

5.2.8 Task Type

We used two different target types in our research: *Single* and *Dual*, and formulated the following secondary research question: "Does the performance of each visual encoding vary with the type of visual search task (*Single* or *Dual* target visual search)?" We evaluated whether the task type also has a significant effect on the performance of the visual encodings.

For *Vizent*, when the task target was *Single*, no significant difference was found between *Task 1* and *Task 2* regarding response time ($p = 0.106$) and accuracy ($p = 0.806$). Not surprisingly, compared to *Single* target tasks, a significant effect of *Dual* target on response time was found ($X^2 = 12.53$, $df = 2$, $p < 0.01$), with the post hoc analysis revealing that a faster response time for *Single* target.

For *Num*, when the task target was *Single*, no significant difference in performance was found between *Task 1* and *Task 2*. Also, no significant effect of *Dual* target was found in response time ($p = 0.31$) and accuracy ($p = 0.19$) compared to *Single* target tasks.

For *Wid–Lig* and *Sat–Tra*, we could not investigate differences between *Single* and *Dual* targets since significant differences in accuracy were found even among *Single* target tasks. We

found that their performance on *Task 2* significantly decreased. This could be in large part due to the transparency level having an adverse effect on the saturation level when *Sat–Tra* was used. Additionally, the lightness level of the edge that interferes with the edge width’s appearance could negatively impact the performance of *Wid–Lig*.

5.3 Experiment 2: Node and Edge Performance Experiment

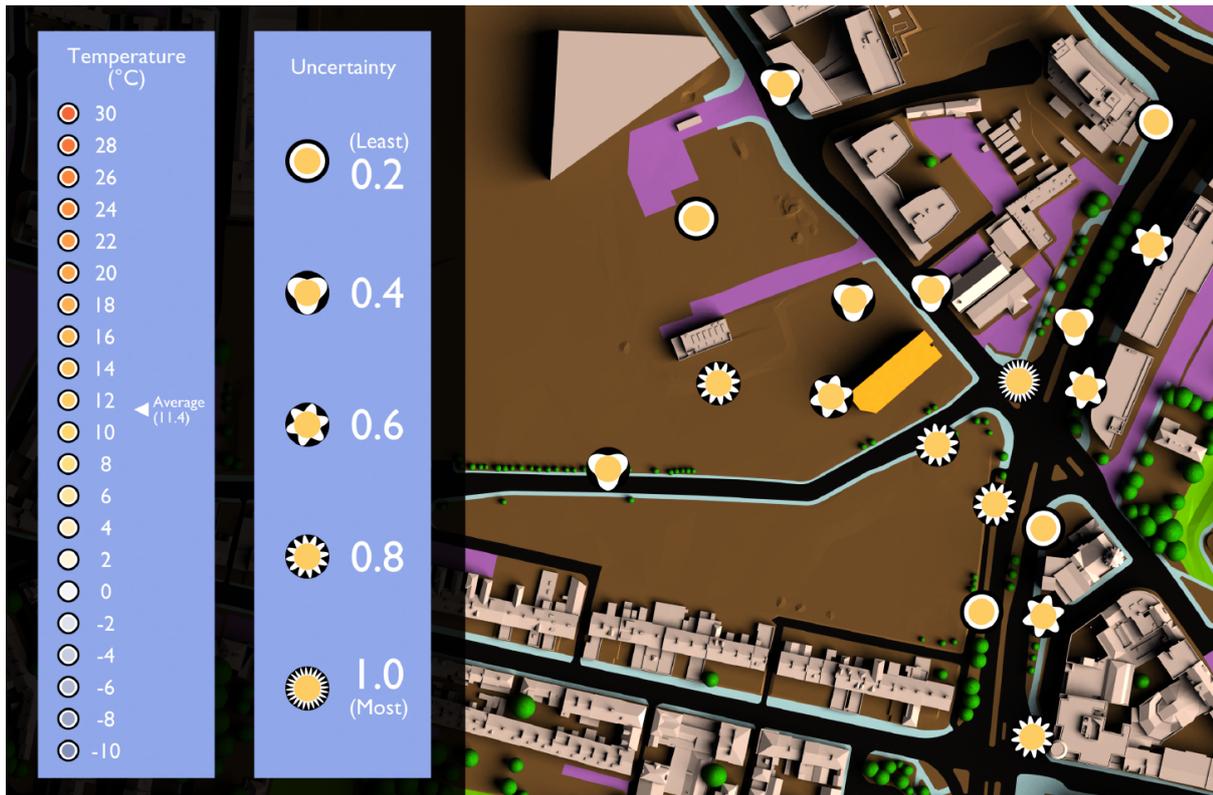


Figure 5.13 Visual entropy glyphs encode hourly mean temperature values using the Met Office colour scale and the variance of those values in the urban digital twin application context [80].

Following the analysis of the initial experiment results, it was determined that our Vizent edge design showed better performance than the visual encodings previously identified from the literature for all tasks, with the exception of the *Num*. Notably, no statistically significant difference was found between the approaches (*Vizent vs Num*) regarding their effectiveness. Consequently, we proceeded to incorporate Visual entropy (Vizent) glyph shapes (see Figure 5.13) as node embeddings. This integration was seamlessly achieved without disrupting the graph layout, as nodes are typically represented as circular entities. The main idea of our approach is to embed Vizent glyph shapes into graph nodes to visualise node-oriented data (bivariate node attributes). The motivation behind our research is the recognition that both nodes and edges can be influenced by uncertainty. By incorporating the representation of uncertainty alongside the primary value, we aim to provide end-users with the ability to make well-informed and efficient decisions.

We evaluated the usability of the Vizent edge design with the Vizent glyphs that were proposed in the study by Holliman et al. [80]. The Vizent glyphs that replace the node representations were used to show bivariate data attached to nodes, as can be seen in Figure 5.14.

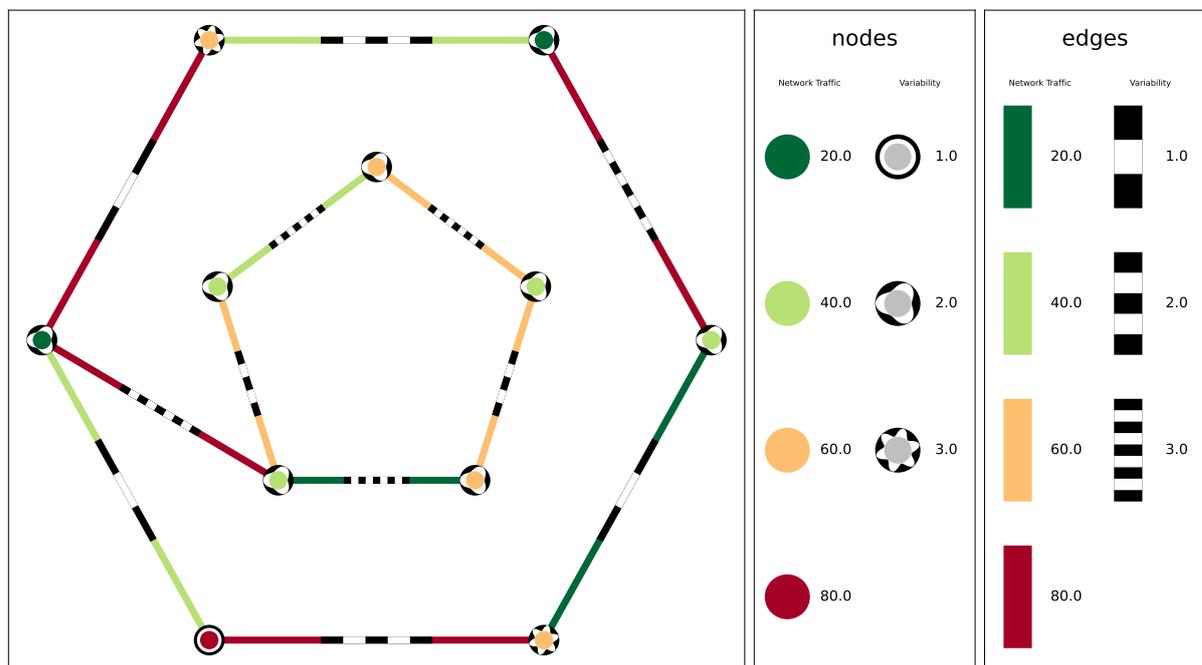


Figure 5.14 Example of the Vizent graph design. Node variables are mapped to Vizent glyph shapes.

A follow-up online experiment with a task-oriented approach was employed following a similar approach to the initial study. The Vizent graph design, see Figure 5.15, was compared against the numerical values (*Num*) visualisation, see Figure 5.16, that had better accuracy than the control group of visual encodings in the first study.

For encoding the edge colour of the Vizent graph, we chose a set of colours that we believed were easy to search for and screened participants for colour vision. There is a lot of literature on colour choice, and in practice, users may choose from many alternative encodings, for example, [153, 45].

Before the experiment, the participants in the online test environment were asked to provide informed consent. The users were informed about the aim of the study before being given the instructions to read. We requested participants use a screen resolution of at least 1920 x 1080 pixels.

We have designed a practice session for each visual graph method to evaluate participant attention to the task. Participants' results were disregarded if respondents did not provide at least one correct answer for each visual approach presented. We deemed these likely to be random responses by the participants as we had no control over the study environment, thus increasing the result's reliability. Additionally, we were careful not to coach participants until they could do the task, and the practice trials were not intended to provide comprehensive training on graph visualisation. Furthermore, to maintain the validity of the experimental design and eliminate the possibility of any prior experience with Vizent approaches, participants who participated in

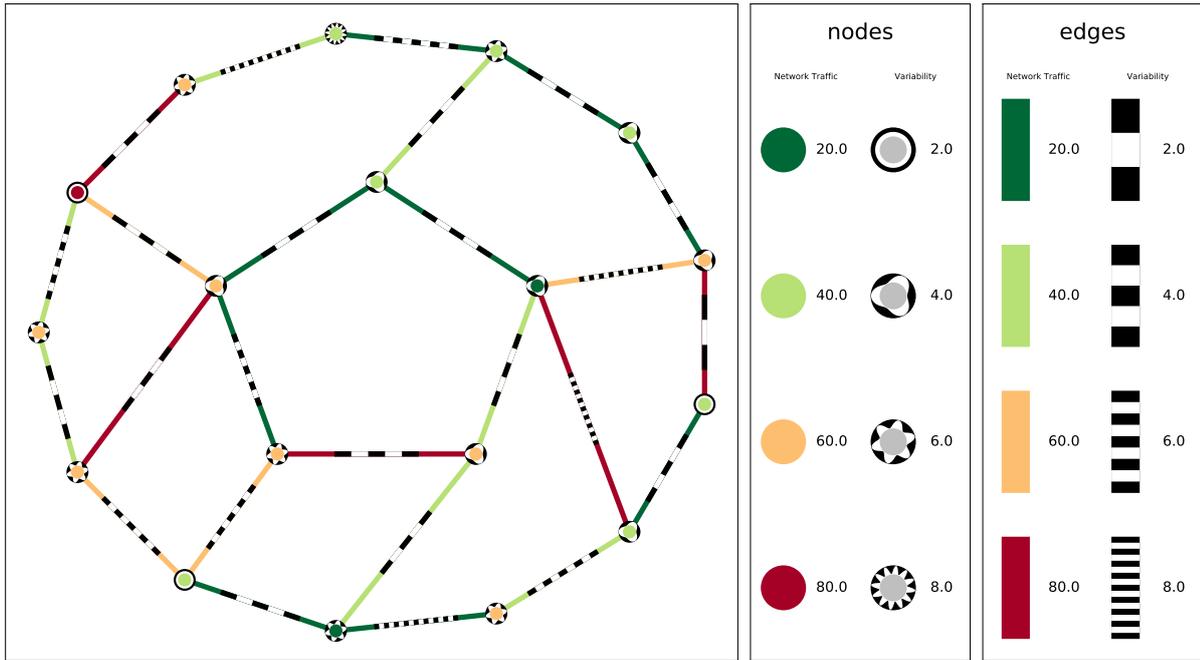


Figure 5.15 Illustration of the Vizent graph design. Four encoding levels for network traffic value and its variability were assigned to Vizent edges and Vizent glyphs [80]. Each edge and node represent the combinations of a network traffic level (20 to 80) and its variability level (2 to 8).

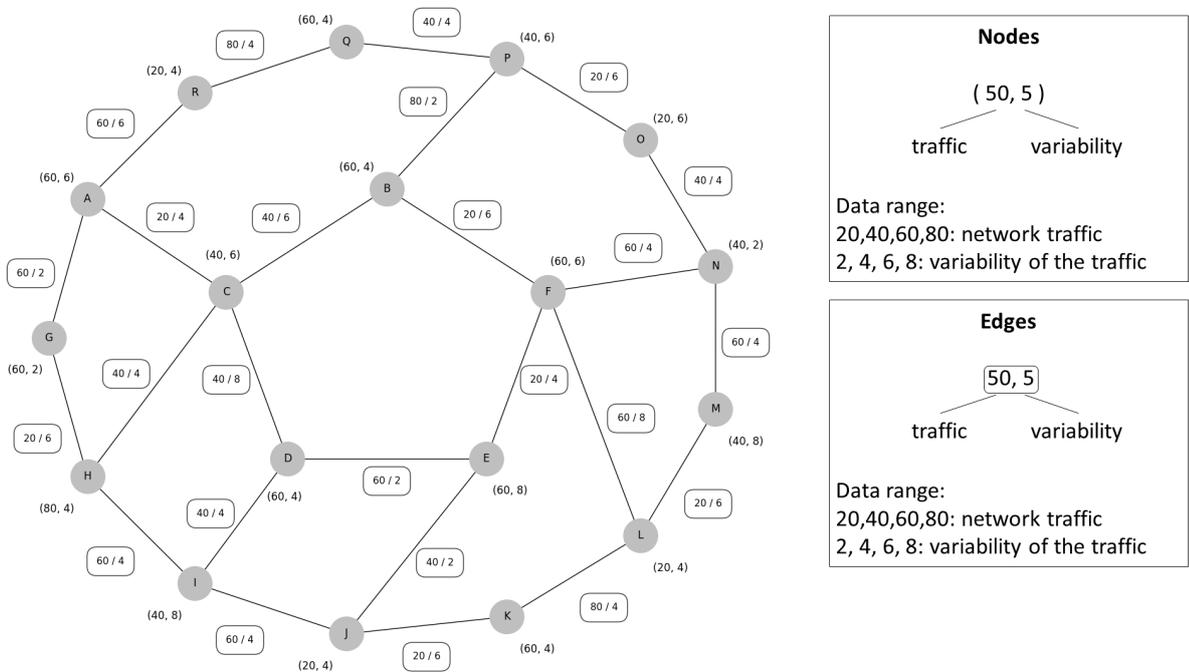


Figure 5.16 Illustration of *Num* visualisation. Four encoding levels for network traffic value and its variability were represented by numerical values. Each edge and node represent the combinations of a network traffic level (20 to 80) and its variability level (2 to 8).

experiment 1 were intentionally prevented from participating in experiment 2 using the Prolific system.

This study makes use of graph visualisation of randomly created datasets with 18 nodes and 25 edges, as in [71, 13]. We utilise graphs without edge crossing to improve the readability of both representations. Each visualisation approach comprises ten distinct graph representations, each with its own layout, see Appendix B. Each participant was instructed to perform a predefined task on both *Num* and the Vized graphs. They are required to identify the edge with the highest network traffic value while also having the lowest variability in that traffic, the same as *Task 3* in the initial study. Participants only click on the middle of the edge if they discover the targeted edge. Each session lasted around twelve minutes. Although we requested participants to complete the task as quickly as possible, we did not put time constraints on the task. We recorded response times and the correctness of the answer for each of the 20 trials in the experiment.

5.3.1 Hypotheses

We hypothesised that two graph visualisations are equally effective, and that participants would perform comparably well in terms of response time and accuracy when completing the predefined task. The following null hypotheses were therefore tested:

H1 There is no difference in the mean response time between the two visualisation approaches when completing the task.

H2 The Vized graph design performs equally well compared to *Num* in terms of accuracy.

We formulated hypotheses about the performance of each graph representation and compared them to the user study's findings.

5.3.2 Experimental Results

We discovered that neither accuracy nor response time was normally distributed following the Shapiro-Wilk test with a significant level of $\alpha = 0.05$. Thus, we conducted a Wilcoxon signed rank test with a significance level set $p < 0.05$ to compare response times and accuracy in the Vized graphs with *Num* for our not normally distributed data.

The Wilcoxon signed-rank test revealed that mean response time in seconds was significantly less when the Vized graphs were used ($\bar{x} = 3.415$, $n = 24$) compared to *Num* ($\bar{x} = 7.192$, $n = 24$), $Z = -5.15$, $p < 0.001$, with a large effect size ($r = .74$). These results rejected the first hypothesis **H1**, which claimed that there would be no significance between approaches in terms of response time.

Regarding accuracy, following the test, we can conclude that there is no statistical evidence to imply that the numerical values and the Vized graphs lead to significantly different correct responses ($Z = -1.365$, $p = 0.172$). The mean score for the accuracy percentage of the numerical values is 87.9 compared to 90.8 for the Vized graphs. Since the two-tailed p-value ($p = 0.172$) is bigger than 0.05, the accuracy performance of the participants on the two graph visualisations is identical and cannot be rejected at the 5 percent significance level, proving that the second hypothesis **H2** holds.

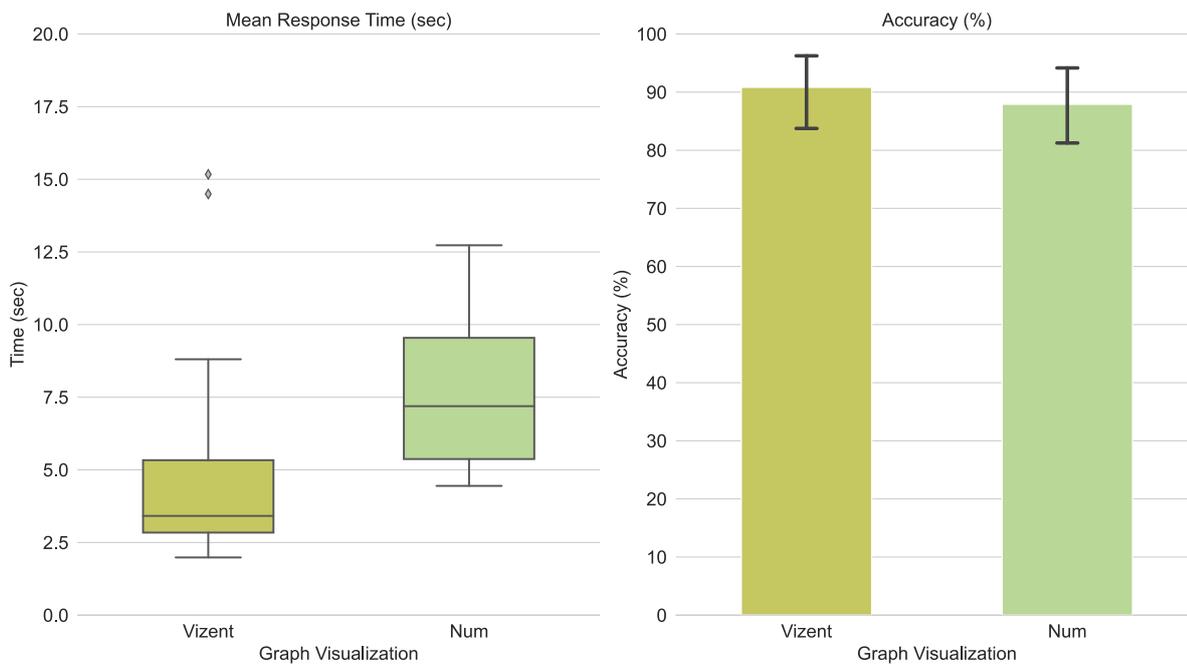


Figure 5.17 Showing the follow-up experiment results. Mean response times in seconds (left side, as box plots) and accuracy (right side, as bars representing means and error bars 95% confidence intervals) classified by Vizent and *Num* graph representations, (n=24).

As demonstrated in Figure 5.17, participants spent a significant amount of time looking for numbers. In comparison to *Num*, which took an average of 84.08 seconds to complete all trials with an overall accuracy of 87.9 percent, the Vizent graphs took an average of 49.29 seconds to complete all trials with an overall accuracy of 90.8 percent. This allowed the participants to quickly gain an accurate visual comprehension of the ordered bivariate values that may be difficult to discern using traditional methods.

Compared to the Vizent graphs, participants took longer to locate the correct answer for numerical values visualisations, which can be explained by addressing established theories in perception and cognition, as proposed by Bertin [23]. Numerical representations necessitate sequential processing, where each value must be read and interpreted individually, which demands high cognitive effort and time. Moreover, the human eye has limited resolution, and when numerical values become too small, densely packed, or complex, they challenge the viewer's ability to discern them quickly. On the other hand, Vizent design uses visual variables, specifically colour and texture, to encode information, taking advantage of the brain's capacity to process visual information simultaneously. This method utilises pre-attentive processing, where certain visual cues are instantly recognised without requiring conscious attention. The rapid processing enables users to comprehend information quickly and effectively, hence reducing the time needed for comprehension. As a result, the Vizent graph should be considered for developing a time-critical bivariate node-link visualisation. We believe the Vizent graph design is superior and will become more common because of its ease of understanding and information presentation.

5.3.3 Participants' comments on the experiment

At the end of the experiment, we asked participants to share their thoughts on the experiment, especially if any of the representations they saw seemed particularly easy or hard to understand. Some of the participants provided their thoughts regarding the experiment as follows:

- *Once I got the hang of it, I felt the coloured lines and thickness of the black/white lines helped the time needed to search for the correct line.*
- *Once I worked out what to do, that was fun! It was easier to find the pattern/colour-coded edges, I think.*
- *I probably found the numbers took slightly longer to find because the visual colours stood out more.*
- *I thought the experiment was interesting and engaging. No issues.*
- *It was odd how quickly the concept became clear because it seemed very obtuse to start.*
- *Coloured representations appeared easier.*
- *The representation of nodes and lines in colours and coloured schemes was easier to understand.*
- *The experiment was tricky, especially at first, but it all went well after the trial.*
- *The traffic light-coloured ones were easier.*
- *At first, I found it quite hard to understand where to click, but once I got the hang of it, it was relatively straightforward.*

In general, the feedback on the experiment was positive. While the experiment seemed interesting but difficult at first, participants were positively engaged and able to complete the experimental task successfully after overcoming the initial phase of the trial. The participants found the numbers version slightly more difficult and required more time. The consensus was that the Vizent graph stood out more and facilitated faster search times, thus superior to the numerical values representations. The comments suggest the experiment offered a compelling test of visual search skills, with participants finding the Vizent graph provided easier pattern recognition and understanding.

5.4 Discussion and Conclusion

This chapter explored the development of a novel visualisation approach to facilitate the visualisation of bivariate network design and the operation of visual search tasks on node-link diagrams. We conducted two task-based usability experiments to evaluate the performances of our proposed new node-link model, named visual entropy (Vizent) graph. For the initial study, the Vizent design was compared to the three effective visual encodings (*Num*, *Wid–Lig*, and *Sat–Tra*) in

the literature to depict bivariate variables of undirected graphs. This included graphs of varying sizes and tasks of varying degrees of difficulty. For the second experiment, the Vizent graph was evaluated against *Num*, which had better accuracy than the control group of visual encodings in the first study.

We found that the efficiency and effectiveness of the visual encodings evaluated vary depending on the task complexity and graph complexity in edge size. The participants achieved higher accuracy of their responses using Vizent and *Num*; however, both *Wid–Lig* and *Sat–Tra* did not show equal performance for all tasks. Moreover, our experiment results show that varying graph size did not affect Vizent; conversely, participants' response time increased significantly on *Num* when the graph size exceeds 15. Furthermore, it does not appear that the size variation had a major effect on the response time and accuracy of *Wid–Lig* and *Sat–Tra*.

Many studies evaluating the efficacy of visual channels have been published in the literature of perceptual studies; the majority agree that pop-out effects aid users in locating information; it is also important to note that the strength of colour in an optical channel relative to the other channels (e.g., width, number) is typically much more noticeable to represent the ordering of value [149, 155]. Additionally, while visual channels selected to encode uncertainty information (e.g. lightness, transparency) introduced additional complexity in the visual representation, the design of the Vizent edge glyph assisted the participants in making an accurate judgment.

Visual channels employing hue and granularity for representing bivariate data have some potential weaknesses which the Vizent design aims to address. They combine hue and granularity within the same region, and the levels of granularity change the shape of the marks, which could affect colour perception. Additionally, any use of excessively light colours, or colours which otherwise lack contrast with the background, could impact the viewer's ability to perceive the granularity correctly [176]. On the contrary, in the Vizent design, the coloured regions maintain a uniform shape and width, while the striped regions maintain a high level of contrast by using black and white stripes.

The experiment results demonstrate that Vizent consistently performs well across various tasks and graph complexities regarding edge size, indicating its robustness and scalability. The performance of Vizent's design could be attributed to its adherence to Gestalt principles in visualisation. These principles enable faster and more accurate interpretation by exploiting the capabilities of human visual processing. Munzner's research [133] demonstrates that the expressiveness principle greatly improves cognitive understanding by organising information in a way that can be easily perceived rather than relying on memory. This also helps users to focus on the most relevant data aspects. The design of the Vizent model aligns with these concepts by emphasising visual order and reducing mental effort, resulting in noticeable enhancements in user task performance.

Our results are promising for the general applicability of the Vizent model to graphs of various sizes and types in domains where it is challenging to handle the volume and complexity of data using traditional visualisation tools. While the results demonstrate potential, further investigation is necessary to ascertain whether they can be implemented in real-world scenarios involving complex and large graphs. Furthermore, it is crucial to acknowledge that although

Vizent exhibited better performance in controlled environments, practical implementations might introduce factors that were not considered in the specified parameters.

In conclusion, this study introduced a novel node-link visual model designed for representing bivariate data to add to the choice of representations available to end users. We believe our findings are sufficiently encouraging to support using the Vizent graphs for implementation.

5.5 Limitations

The results of this experiment need to be evaluated in the context of the experiment's parameters. Although we attempted to be as diverse as possible in our selection of participants, graph sizes, and tasks, it is hard to account for every possibility.

The chosen colour scale was not evaluated for colour-blind users in the follow-up experiment because none of our test subjects with a colour vision deficiency. On the other hand, the tested colour scheme is not suited for colour-blind users; any colour combinations involving red, green, or any shades in between can create disorientation for those with this condition. However, it is easy to choose an alternative scheme for colour-deficient viewers matched to their deficiency type.

The produced graphs were limited to scale-free graphs, and a specific layout approach (shell layout) was employed. Other graph layouts, such as a force-directed layout, may influence the effectiveness of task performance. While we expect our findings to be applicable to all node-link diagrams, additional research might be necessary to confirm this.

Due to time and access constraints, we were able to conduct online experiments with participants to test the performance of our proposed node-link approaches because of the pandemic restrictions. The utilisation of crowd-sourced evaluation offers the advantage of substantial sample numbers, although it is not without its associated challenges [9]. In order to enhance the reliability of our findings, we established our filtering criteria by considering the replies deemed successful in the attention checks administered prior to the study. The results of participants were excluded from the analysis if they failed to offer at least one accurate response for each visual approach provided. For the second experiment, 50 distinct individuals were initially recruited to participate in our online experiment. However, only 48 of them completed the experiment. Nevertheless, it is noteworthy that 50% of the participants met the predetermined criteria for inclusion in our analysis.

5.6 Future Works

The usability studies have shown convincing results; however, as mentioned earlier, the sizes of the tested graph are not excessively enormous, which aligns with findings reported in existing research [71, 13]. Additionally, we selected an edge size of 25 since it was deemed to be the most equitable choice for all visual encodings, taking into account the absence of interactivity in the experiment. We believe, on the basis of our experience, that the results of the experiment would be applicable even when considering network sizes above 25. Nevertheless, it remains

an open question how each visual encoding tested would scale when the graph size is beyond this threshold. In future work, it would be valuable to conduct a more detailed evaluation to investigate the impact of different graph sizes, specifically those with 50, 75, and 100 edges, on the scalability and efficiency of our proposed visual model within a predetermined experimental setting. This work will provide valuable insights into the performance of visual models under varying edge conditions.

The proposed approaches have not been tested on multigraphs, and future studies will focus on conducting empirical research on employing Vizent graphs on curved edges. In real-world case scenarios, certain nodes may have more than one edge and curved edges could be employed to prevent edge crossings. Additionally, it might be insufficient for single-line connections to capture the complex information inside a network existing in the physical world. We noted it for future work, and it would need to be verified experimentally to see how well our design works on curved edges. Therefore, it is part of our future agenda to conduct contextual research to observe how effectively the Vizent graph functions in more realistic environments of the target users.

Although the primary focus of our research has been to develop effective static two-dimensional visual representations for representing bivariate data, extending our novel design into dynamic 3D environments and incorporating multivariate data would be an interesting direction for future work. Initially, the Vizent graph could be improved to effectively represent multivariate data, hence expanding its applicability. This could involve implementing visual encoding methods, such as varying edge thickness, to encode additional variables without cluttering the visualisation. Moreover, future research could focus on transitioning from static 2D visual representation to dynamic virtual reality (VR) or augmented reality (AR) settings for handling large multivariate graphs in real-world scenarios. Due to the graph's manageable size in terms of screen space and the limited graphical space of the edges in a 2D visualisation, integrating the Vizent graph with interactive technologies could improve multivariate graph visualisation. This could allow users to explore graph structures from multiple angles and interpret the multivariate information represented in the Vizent graph when projecting onto 3D.

Chapter 6

Conclusion

Contents

6.1	Summary of Contributions	122
6.2	Answers to research questions	123
6.3	Future Work	124
6.3.1	Generalisation	125

This final chapter discusses the thesis results, including the main findings and contributions from the preceding chapters (see Section 6.1). It also provides concise responses to the research questions proposed (see Section 6.2), thereby integrating the key results and insights acquired throughout the development of this research. Finally, we discuss possible directions for future work (see Section 6.3).

We present the concept of edge variance in the context of edge uncertainty and a summarisation approach to how to get the maximum information while summarising large non-hierarchical multi(di)graphs with non-negative edge weights in Chapter 3. Given the limitation on the number of user-defined edges, the approach reduces the graph size by merging multiple edges into a single representative edge called a *superedge* while retaining concise summaries highlighting the most variable and information-rich connections. The resulting summary allows the end-users to focus on more “interesting” edges, defined in terms of edge variance while displaying much more information given the same number of edges from an information theoretical point of view.

Our summary algorithm currently does not consider the possibility that certain nodes may be more interesting than others; this was acknowledged as a limitation, and future work will concentrate on the variability of both edges and nodes in real-world scenarios where ignoring node variables may result in the omission of interesting patterns. An additional limitation is that the dataset we evaluated encompassed a maximum of 1500 distinct edges; additional investigation is also required to ascertain the scalability of our algorithm.

Despite the pervasive nature of uncertainty, it is frequently not well conveyed to the user or — potentially much more problematic — disregarded during the initial phases of the visualisation pipeline. On the other hand, various visual channels are available to represent two values simultaneously in the context of node-link diagrams. However, these approaches are not generally effective in reflecting an uncertainty value due to the inference between visual channels. These visualisations, in turn, may lead to flawed decision-making by end-users when reading bivariate graph visualisations due to the potential for misinterpretation if they are not meticulously developed. This challenge is nontrivial due to the limitations of edges, including their size and the restricted capacity of individual channels.

As previously discussed in the thesis, this thesis is primarily concerned with the issue of representing uncertainty in (static) node-link diagrams inside a two-dimensional visual space to enhance comprehension for the general public. We developed novel node-link visualisation techniques — *Visual entropy* graphs while overcoming some of the challenges of bivariate graph visualisation approaches. Furthermore, it is important to note that the datasets may contain missing uncertainty values, which must be managed cautiously. Therefore, we meticulously devised a distinct design to represent missing values visually.

Visual search-based tasks are an essential activity in the real world. In our research questions, we explore how well the visual encodings perform the judgement of min and max values. This type of visual search task is common for numerous visualisation objectives, such as anomalies and outlier detection. To evaluate our visualisation method, we conducted two empirical task-based usability studies that compared how effective and efficient the Vizent approach is with respect to the existing bivariate attribute visualisations, namely Width–Lightness, Saturation–Transparency, as well as Numerical values.

The first usability study yielded the following results: Participants demonstrated increased response accuracy when utilising *Vizent* and *Num*. However, it should be noted that *Wid–Lig* and *Sat–Tra* did not exhibit consistent performance across all tasks. Additionally, the variation in graph size did not have a discernible impact on the effectiveness and efficiency of *Vizent*. No significant difference was observed in the performance of *Vizent* across different graph sizes, ranging from 5 to 25 edges. The study also examines the potential complexity of task targets, namely *Single* and *Dual* targets. In the case of *Vizent*, when the task aim was set to *Single*, there was no statistically significant difference between *Task 1* and *Task 2* regarding response time ($p = 0.106$) and accuracy ($p = 0.806$). Consistent with expectations, the present study observed a notable impact of *Dual* target conditions on response time ($X^2 = 12.53$, $df = 2$, $p < 0.01$). Further analysis indicated that *Single*-target conditions resulted in a comparatively quicker response time for *Vizent* design.

Literature has shown, parallel to our findings, that the user's performance is enhanced by visual channels that are perceptually orderable concerning our new edge design [39]. Moreover, the colour (hue), encoded to represent the network traffic for *Vizent*, provided a method for quickly determining the targeted edge compared to the control group of visual encodings. In contrast, lightness and transparency likely harmed the performance, significantly reducing the effectiveness (accuracy) of the visual encodings. For *Wid–Lig*, the lightness could reduce the discriminability of the edge mark (width level). For *Sat–Tra*, using saturation with varying levels of transparency may have added distracting complexity to the visualisation due to the interference and interplay between the two distinct visual variables.

Consistent with previous research, the outcomes of Experiment 1 (Section 5.2) demonstrated that effective visualisation of bivariate attributes on edges within the framework of node-link diagrams is not straightforward. A greater understanding of the interplay between multiple channels is necessary to facilitate effective visualisation [176]. The findings presented in the first experiment proved that our design can be reliably utilised to understand the intended values. Our design facilitates a correct perception of differences through a sequence of *Vizent* edges that are visually ordered and designed so that each sequence member of *Vizent* edge is distinct from the previous one in a doubled pattern frequency.

In the second experiment, based on the quantitative results and qualitative feedback, it can be concluded that our novel node-link visual approach efficiently facilitates the simultaneous comparison of dual attributes through a static representation. We also concluded that the *Vizent* graph provides an efficient and effective approach for communicating bivariate data, especially when users need to quickly comprehend the uncertainty of the data to make crucial judgements under tight time constraints. Our proposed visual graph reduces the necessity of learning and memorising the coding scheme linked to the *Vizent* glyphs.

Our study results provide practical guidance for researchers seeking an effective visualisation approach to encode bivariate edge information, particularly data value and uncertainty information, in the context of node-link diagrams. Drawing from the findings of our experiments, researchers can explore the performance of the alternative visual encodings tested in our user studies and how *Vizent* design works well against them for bivariate representations. By adapting our *Vizent* graph design, experts and/or the general public can create visualisations that

maintain consistent perceptual differences between visual cues, enabling effective information communication. In conclusion, we believe that our studies can serve as a basis for more informed information, particularly uncertainty encoding on bivariate edge attributes.

6.1 Summary of Contributions

Each novel contribution was published in the *Visual Informatics* journal, under the title: **Visualizing ordered bivariate data on node-link diagrams** [1]. The most significant research contributions made by this thesis are as follows:

- *A novel edge design for effectively representing both primary and secondary values, such as uncertainty, on the edges simultaneously in the context of node-link diagrams. Empirically demonstrating that different edge glyphs have a perceived order through pairwise testing.*

As previously discussed, most existing visualisation methods are not specifically designed to handle abstract data. Moreover, their primary focus revolves around the representation of a single value or uncertainty value. In relationships, methods for communicating uncertainty are lacking. Relationships within the data hold significant value in numerous decision-making scenarios. Thus, it is needed to carefully and accurately exhibit uncertainty in relationships. One overarching difficulty is the creation of visual metaphors that establish an intuitive cognitive connection between uncertainties and the associated data, ensuring that users do not fail to perceive this relationship [69].

Chapter 4 presents the novel edge design to represent ordered categorical information or quantised numerical data on interval values and the evaluation of the proposed glyphs to confirm the perceptual effectiveness of the Vizent edge glyphs. Vizent edges are designed so that the user can reliably understand the desired ordered values. This is achieved by a visually ordered sequence of Vizent edges, where each sequence member is distinct from the previous one regarding doubling pattern frequency. Our empirical evidence suggests that the participants perceive the visual entropy edge glyphs in an order. Because of this perceptual effect, the dependence on legends that show encoding meanings is greatly diminished, often leading to a substantial improvement in overall performance [116].

- *Performing two task-based usability studies to demonstrate the efficiency and effectiveness of our approach for visualising bivariate networks in the context of static node-link diagrams.*

User studies are of great significance as they aid the visualisation community in selecting appropriate techniques for their tasks and are instrumental in assessing the effectiveness of various uncertainty visualisation techniques [97]. We presented two experiments investigating the efficiency and effectiveness of Vizent design in the context graphs drawn as node-link diagrams in Chapter 5.

- *Comparing the Vizent design against three visual encodings selected from the literature on various graphs ranging in complexity from 5 to 25 edges for three different tasks.*

We formulated the following research question: “Does the Vizent design perform equally well or better against the three visual encodings for all tasks in terms of response time and accuracy?” as well as two secondary questions with respect to task complexity and graph complexity. To address these questions, we conducted a controlled experiment through Prolific in which participants were asked to complete graph-related tasks. We tested two task types and five different edge size levels. Our study findings show that the performance of each task varies significantly depending on the choice of visual encoding tested. Additionally, we observed that the complexity of the graph does not significantly affect the performance of each visual encoding, with the exception of Numerical values (*Num*) visualisation.

- *A novel node-link visual model for effectively representing both primary and secondary values, such as uncertainty, on the edges simultaneously.*

Finally, this thesis presents the Vizent graph, which helps with the general-purpose visualisation of bivariate data for graph exploration and complete consideration of the uncertainty linkage of edges and/or nodes through the effective graphical metaphors of the proposed 2D node-link visual model. This allows viewers to acquire an intuitive sense of the secondary value of the underlying data. Furthermore, it can serve as a valuable tool for enhancing existing visualisation methods and functioning as a stand-alone visualisation approach.

6.2 Answers to research questions

Upon reviewing succinct explanations of our research contributions, including the design process, integration of novel methodologies, and validation through user studies, we are now able to delve into the research questions that guided our investigation and furnish well-grounded and empirical responses:

- **Q:** *Exploring a scalable approach for visualising complex graph-based data, including measures and variance of measures?*

A: The first research question focuses on a scalable approach integrating their variances as a measure of uncertainty into the summarisation process. We define the concept of edge variance, which measures the extent of variability or uncertainty in the weights or attributes associated with edges of a graph. We proposed a maximum entropy-based graph summarisation approach that identifies edges with the most informative relationships given the user-defined number of edges.

- **Q:** *How to design a novel entropy-based bivariate representation of networks?*

A: The second research question focuses on developing a novel 2D bivariate visualisation method to provide users with an alternate static representation that is easily understandable in addition to the conventional node-link diagrams to assist users in making improved and well-informed judgements. Therefore, we introduced — a *Visual Entropy* (Vizent)

graph — that enhances conventional node-link diagrams. A set of criteria facilitated the concentration of our research efforts on addressing tangible issues that contribute to attaining the objectives. We showed how they can be utilised to represent bivariate data (see Section 4.5.1). We provided practical impacts for real-world challenges involving two variables in the context of uncertainty.

- **Q:** *How do entropy-based representation networks compare in visual search performances to existing approaches?*

A: We designed our user studies as a quantitative controlled task-based evaluation involving two observed dependent variables: response time and accuracy, where participants perform visual search tasks while ensuring that tasks represent real-world scenarios.

The initial experiment compared the Vizent approach with several visualisation encodings, while the second experiment focused on comparing it with the numerical values visualisations. The purpose of this empirical evaluation was to evaluate the effectiveness and efficiency of these encodings in tasks related to visual search on bivariate graphs. Our evaluation showed that the Vizent design produced promising results empirically for given tasks. Furthermore, our qualitative and quantitative observations from the second experiment indicated that the Vizent graph effectively represents the features of bivariate data and improves response time and accuracy.

In summary, this thesis provides greater insight into understanding and representing bivariate graphs. It opens up new possibilities for applying these data representations to real-world scenarios. The novel methodologies and insights presented in this study advance the field and lay the groundwork for future research in information and graph visualisation. The findings of this thesis point toward some intriguing future directions.

6.3 Future Work

This research has several possible future research directions. Graph summarisation approaches have been widely utilised in several domains, such as traffic, social, and citation networks. On the other hand, the application of graph summarisation techniques in intelligent military and defence systems has been limited despite its potential as a valuable tool in this field. A node-link diagram could be used to show this idea using graph-structured data from Internet of Things (IoT) devices [47], such as security cameras and motion sensors, based on the generated network traffic. Despite this, the large amount of data makes it hard for commanders to find critical information about possible anomalies and changes in the communication network caused by IoT devices while also staying aware of what is happening. Therefore, applying the proposed graph summary approach could address these challenges and help uncover hidden patterns and detect anomalies. Since commanders often lack expertise in interpreting variance measures and require prompt decision-making, our research aims to support them by integrating the suggested graph summarisation algorithm with our novel graph visualisation design. This combination could facilitate gaining profound insights and allow for a quick overview of the information.

This future work will contribute to advancing graph-based data analysis and developing robust solutions for practical application.

6.3.1 Generalisation

In modern times, where the emphasis on data-driven decision-making is of utmost importance, not only do data representation and uncertainty play a crucial role, but also making data accessible and understandable to the general public is an important task. Our approach introduces a novel and effective solution that outperforms traditional techniques of visualising bivariate data in the presence of uncertainty. While initially developed for the specific domain of uncertainty visualisation, this research can be successfully expanded and generalised in various directions involving two dependent data variables because of its visual simplicity and adaptability.

Graph drawing and cartography intersect when the visualisation of *geometric networks*, which consist of elements with geographic locations or geospatial aspects, is required [182]. These networks include street layouts, underground metro, river, and cable networks. Our visualisation method can demonstrate how well it performs at the intersection of graph drawing and cartography for the visual analysis of bivariate geometric networks.

Our visualisation approach is applicable to a diverse range of network data sets and any network scenarios. Geometric networks are integral to many real-world scenarios. For example, in the context of a city street map, the objective is to delineate traffic congestion and air quality at various urban locations by adopting bivariate data visualisation techniques. Vized design can be employed to convey both traffic congestion and air quality data. With the Vized edge glyphs depicting traffic congestion levels, the colour hue can represent different air quality levels along the road segments, which could offer users a clear understanding of the environmental conditions across the city's road network.

An additional illustration of the potential application of our novel visualisation approach involves interacting with Transport for London (TfL) rail data in the context of the London Underground map. For instance, we could enhance metro stations by introducing Vized glyphs on nodes to convey passenger traffic and its variation at each station. Furthermore, the lines could incorporate Vized edges, in which colour hue represents train capacity occupancy and the Vized edge glyphs represent the time-dependent variations in train frequency, train intervals or service disruptions. This has the potential to assist individuals in organising their journeys more efficiently and adapting to changing conditions.

Generalising the Vized graph for health data systems could improve the visualisation and interpretation of bivariate data, particularly in the context of uncertainty visualisation. Due to its ability to encode bivariate data, the Vized graph design would be applicable in visualising the spread of contagious diseases and their uncertainties across various geographic locations or populations. Our design could map the transmission of diseases in a way that highlights the intensity of outbreaks and the uncertainties in transmission rates, making it a valuable tool in public health management. Through the Vized graphs, health authorities may be able to rapidly identify areas where the transmission rate is high and highly variable, thereby locating the areas where potentially more hazardous and unpredictability of the disease's spread. This strategy

could support ongoing monitoring and management of disease spread and assist in immediate response efforts when additional research is required to determine the underlying causes of disease spread variability. Moreover, the Vizent design's ability to display these complex data relationships in a clear and intuitive format makes it an invaluable tool for communicating the dynamic and often uncertain nature of disease outbreaks to non-expert stakeholders, thereby enhancing community engagement and compliance with health measures.

Bibliography

- [1] Akbulut, O., McLaughlin, L., Xin, T., Forshaw, M., and Holliman, N. S. (2023). Visualizing ordered bivariate data on node-link diagrams. *Visual Informatics*, 7(3):22–36.
- [2] Alencar, A. B., Börner, K., Paulovich, F. V., and de Oliveira, M. C. F. (2012). Time-aware visualization of document collections. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 997–1004.
- [3] Alper, B., Hollerer, T., Kuchera-Morin, J., and Forbes, A. (2011). Stereoscopic highlighting: 2d graph visualization on stereo displays. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2325–2333.
- [4] Amar, R., Eagan, J., and Stasko, J. (2005). Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization*, pages 111–117. IEEE.
- [5] Andreyev, A. (2014). Introducing data center fabric, the next-generation facebook data center network.
- [6] Archambault, D., Abello, J., Kennedy, J., Kobourov, S., Ma, K.-L., Miksch, S., Muelder, C., and Telea, A. C. (2014). Temporal multivariate networks. In Kerren, A., Purchase, H. C., and Ward, M. O., editors, *Multivariate Network Visualization*, chapter 8, pages 151–174. Springer International Publishing.
- [7] Archambault, D., Munzner, T., and Auber, D. (2007a). Topolayout: Multilevel graph layout by topological features. *IEEE transactions on visualization and computer graphics*, 13(2):305–317.
- [8] Archambault, D., Munzner, T., and Auber, D. (2007b). Topolayout: Multilevel graph layout by topological features. *IEEE transactions on visualization and computer graphics*, 13(2):305–317.
- [9] Archambault, D., Purchase, H., and Hoßfeld, T. (2017). *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Springer.
- [10] Armstrong, R. A. (2014). When to use the bonferroni correction. *Ophthalmic and physiological optics*, 34(5):502–508.
- [11] Bach, B., Pietriga, E., and Fekete, J.-D. (2014). Graphdiaries: Animated transitions and temporal navigation for dynamic networks. *IEEE Trans. Vis. and Comput. Graph.*, 20(5):740–754.
- [12] Bach, B., Riche, N. H., Hurter, C., Marriott, K., and Dwyer, T. (2016). Towards unambiguous edge bundling: Investigating confluent drawings for network visualization. *IEEE transactions on visualization and computer graphics*, 23(1):541–550.
- [13] Bae, J., Ventocilla, E., Riveiro, M., Helldin, T., and Falkman, G. (2017). Evaluating multi-attributes on cause and effect relationship visualization. In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - IVAPP, (VISIGRAPP 2017)*, page 64–74.

- [14] Ball, W. W. R. (1892). *Mathematical recreations and problems of past and present times*. Macmillan and Company.
- [15] Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. *Proceedings of the international AAAI conference on web and social media*, 3(1):361–362.
- [16] Batagelj, V., Brandenburg, F. J., Didimo, W., Liotta, G., Palladino, P., and Patrignani, M. (2010). Visual analysis of large graphs using (x, y)-clustering and hybrid visualizations. *IEEE transactions on visualization and computer graphics*, 17(11):1587–1598.
- [17] Battista, G. D., Eades, P., Tamassia, R., and Tollis, I. G. (1998). *Graph drawing: algorithms for the visualization of graphs*. Prentice Hall PTR.
- [18] Beck, F., Burch, M., and Diehl, S. (2009). Towards an aesthetic dimensions framework for dynamic graph visualisations. In *2009 13th international conference information visualisation*, pages 592–597. IEEE.
- [19] Beck, F., Burch, M., Diehl, S., and Weiskopf, D. (2017). A taxonomy and survey of dynamic graph visualization. *Computer graphics forum*, 36(1):133–159.
- [20] Beck, F., Burch, M., Vehlow, C., Diehl, S., and Weiskopf, D. (2012). Rapid serial visual presentation in dynamic graph visualization. In *Proceedings of the 2012 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 185–192. IEEE.
- [21] Becker, R. A., Eick, S. G., and Wilks, A. R. (1995). Visualizing network data. *IEEE Transactions on visualization and computer graphics*, 1(1):16–28.
- [22] Belia, S., Fidler, F., Williams, J., and Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, 10(4):389.
- [23] Bertin, J. and Berg, W. J. (1983). *Semiology of graphics : diagrams, networks, maps*. University of Wisconsin press Madison.
- [24] Bezerianos, A., Chevalier, F., Dragicevic, P., Elmqvist, N., and Fekete, J.-D. (2010). Graphdice: A system for exploring multivariate social networks. *Computer graphics forum*, 29(3):863–872.
- [25] Borgo, R., Kehrer, J., Chung, D. H. S., Maguire, E., Laramée, R. S., Hauser, H., Ward, M., and Chen, M. (2013). Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Eurographics State of the Art Reports*, pages 39–63. Eurographics Association.
- [26] Börner, K. and Polley, D. E. (2014). *Visual Insights: A Practical Guide to Making Sense of Data*. The MIT Press.
- [27] Boukhelifa, N., Bezerianos, A., Isenberg, T., and Fekete, J.-D. (2012). Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty. *IEEE Trans. Vis. Comput. Graphics*, 18(12):2769–2778.
- [28] Brodkorb, F., Kuijper, A., Andrienko, G., Andrienko, N., and Von Landesberger, T. (2016). Overview with details for exploring geo-located graphs on maps. *Information Visualization*, 15(3):214–237.
- [29] Brodlie, K., Allendes Osorio, R., and Lopes, A. (2012). A review of uncertainty in data visualization. In *Expanding the Frontiers of Visual Analytics and Visualization*, pages 81–109, London. Springer London.

- [30] Burch, M., Huang, W., Wakefield, M., Purchase, H. C., Weiskopf, D., and Hua, J. (2020). The state of the art in empirical user evaluation of graph visualizations. *IEEE Access*, 9:4173–4198.
- [31] Butler, D. M., Almond, J. C., Bergeron, R. D., Brodlie, K. W., and Haber, R. B. (1993). Visualization reference models. In *Proceedings of the 4th conference on Visualization '93*, pages 337–342.
- [32] Cairns, P. (2019). *Doing better statistics in human-computer interaction*. Cambridge University Press.
- [33] Card, S., Mackinlay, J., and Shneiderman, B. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann, California, USA.
- [34] Card, S. K. and Mackinlay, J. (1997). The structure of the information visualization design space. In *Proceedings of VIZ'97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium*, pages 92–99. IEEE.
- [35] Chen, M. and Floridi, L. (2013). An analysis of information visualisation. *Synthese*, 190:3421–3438.
- [36] Cheong, S.-H. and Si, Y.-W. (2020). Force-directed algorithms for schematic drawings and placement: A survey. *Information Visualization*, 19(1):65–91.
- [37] Chimani, M., Gutwenger, C., Jünger, M., Klau, G. W., Klein, K., and Mutzel, P. (2013). The open graph drawing framework (ogdf). In Tamassia, R., editor, *Handbook of graph drawing and visualization*, chapter 17, pages 543–569. CRC Press.
- [38] Chung, D. H., Legg, P. A., Parry, M. L., Bown, R., Griffiths, I. W., Laramee, R. S., and Chen, M. (2015). Glyph sorting: Interactive visualization for multi-dimensional data. *Information Visualization*, 14(1):76–90.
- [39] Chung, D. H. S., Archambault, D., Borgo, R., Edwards, D. J., Laramee, R. S., and Chen, M. (2016). How ordered is it? on the perceptual orderability of visual channels. *Comput. Graph. Forum*, 35(3):131–140.
- [40] Cleveland, W. S. and McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554.
- [41] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Assoc.
- [42] Cordeil, M., Dwyer, T., Klein, K., Laha, B., Marriott, K., and Thomas, B. H. (2016). Immersive collaborative analysis of network connectivity: Cave-style or head-mounted display? *IEEE transactions on visualization and computer graphics*, 23(1):441–450.
- [43] Correll, M., Moritz, D., and Heer, J. (2018). Value-suppressing uncertainty palettes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18*, page 1–11.
- [44] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley, New York.
- [45] Cramer, F., Shephard, G. E., and Heron, P. J. (2020). The misuse of colour in science communication. *Nature communications*, 11(1):5444.
- [46] Csiszár, I., Shields, P. C., et al. (2004). Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528.

- [47] Cvitić, I., Peraković, D., Periša, M., and Gupta, B. (2021). Ensemble machine learning approach for classification of iot devices in smart home. *International Journal of Machine Learning and Cybernetics*, 12(11):3179–3202.
- [48] Dehmer, M. (2011). Information theory of networks. *Symmetry*, 3(4):767–779.
- [49] Dehmer, M. and Mowshowitz, A. (2011). A history of graph entropy measures. *Information Sciences*, 181(1):57–78.
- [50] Dunne, C. and Shneiderman, B. (2013). Motif simplification: Improving network visualization readability with fan, connector, and clique glyphs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI’13, page 3247–3256.
- [51] Eades, P. and Xuemin, L. (1989). How to draw a directed graph. In *IEEE Workshop on Visual Languages*, pages 13–14. IEEE Computer Society.
- [52] Ellis, G. and Dix, A. (2007). A taxonomy of clutter reduction for information visualisation. *IEEE transactions on visualization and computer graphics*, 13(6):1216–1223.
- [53] Ellson, J., Gansner, E. R., Koutsofios, E., North, S. C., and Woodhull, G. (2004). Graphviz and dynagraph—static and dynamic graph drawing tools. *Graph drawing software*, pages 127–148.
- [54] Euler, L. (1736). *Solutio problematis ad geometriam situs pertinentis*. *Comment. Academiae Sci. I. Petropolitanae*, 8:128–140.
- [55] Fan, W., Li, J., Wang, X., and Wu, Y. (2012). Query preserving graph compression. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD ’12, page 157–168.
- [56] Farrington, N. and Andreyev, A. (2013). Facebook’s data center network architecture. In *2013 Optical Interconnects Conference*, pages 49–50. IEEE.
- [57] Fekete, J.-D., Van Wijk, J. J., Stasko, J. T., and North, C. (2008). The value of information visualization. *Information Visualization: Human-Centered Issues and Perspectives*, pages 1–18.
- [58] Finger, R. and Bisantz, A. M. (2002). Utilizing graphical formats to convey uncertainty in a decision-making task. *Theoretical Issues in Ergonomics Science*, 3(1):1–25.
- [59] Fletcher, C., Huang, W., Arness, D., and Nguyen, Q. V. (2019). The role of working memory capacity in graph reading performance. In *2019 IEEE Pacific Visualization Symposium (PacificVis)*, pages 77–81. IEEE.
- [60] Frishman, Y. and Tal, A. (2008). Online dynamic graph drawing. *IEEE Transactions on Visualization and Computer Graphics*, 14(4):727–740.
- [61] Fruchterman, T. M. and Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164.
- [62] Furnas, G. W. (1986). Generalized fisheye views. *Acm Sigchi Bulletin*, 17(4):16–23.
- [63] Gallager, R. G. (1968). *Information theory and reliable communication*, volume 588. Springer.
- [64] Gansner, E. R., Hu, Y., North, S., and Scheidegger, C. (2011). Multilevel agglomerative edge bundling for visualizing large graphs. In *Proc. IEEE Pacific Vis. Symp.*, pages 187–194. IEEE.
- [65] Gansner, E. R. and Koren, Y. (2006). Improved circular layouts. In *International Symposium on Graph Drawing*, pages 386–398. Springer.

- [66] Ghani, S. and Elmqvist, N. (2011). Improving revisitation in graphs through static spatial features. In *Graphics Interface*, pages 175–182.
- [67] Ghoniem, M., Fekete, J.-D., and Castagliola, P. (2004). A comparison of the readability of graphs using node-link and matrix-based representations. In *Proc. IEEE Symp. Inf. Vis.*, pages 17–24. IEEE.
- [68] Ghoniem, M., Fekete, J.-D., and Castagliola, P. (2005). On the readability of graphs using node-link and matrix-based representations: A controlled experiment and statistical analysis. *Information Visualization*, 4(2):114–135.
- [69] Griethe, H. and Schumann, H. (2005). Visualizing uncertainty for improved decision making. In *Proceedings of the 4th International Conference on Business Informatics Research (BIR 2005)*.
- [70] Gudkov, V. (2016). Generalized entropies of complex and random networks. In Dehmer, M., Emmert-Streib, F., Chen, Z., Li, X., and Shi, Y., editors, *Mathematical Foundations and Applications of Graph Entropy*, chapter 2, pages 41–61. Wiley, New York, NY, USA.
- [71] Guo, H., Huang, J., and Laidlaw, D. H. (2015). Representing uncertainty in graph edges: An evaluation of paired visual variables. *IEEE Transactions on Visualization and Computer Graphics*, 21(10):1173–1186.
- [72] Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T., and Millman, J., editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.
- [73] Hassanlou, N., Shoaran, M., and Thomo, A. (2013). Probabilistic graph summarization. In *Proceedings of the 14th International Conference on Web-Age Information Management, WAIM’13*, page 545–556.
- [74] Hautus, M. J., Macmillan, N. A., and Creelman, C. D. (2021). *Detection theory: A user’s guide*. Routledge.
- [75] Henecka, W. and Roughan, M. (2015). Lossy compression of dynamic, weighted graphs. In *Proceedings of the 2015 3rd International Conference on Future Internet of Things and Cloud, FiCloud’15*, page 427–434.
- [76] Henry, N. and Fekete, J.-D. (2007). Matlink: Enhanced matrix visualization for analyzing social networks. In *Proceedings of the International Conference on Human-Computer-Interaction – Interact*, pages 288–302. Springer.
- [77] Henry, N., Fekete, J.-D., and McGuffin, M. J. (2007). Nodetrix: a hybrid visualization of social networks. *IEEE transactions on visualization and computer graphics*, 13(6):1302–1309.
- [78] Herman, I., Melançon, G., and Marshall, M. S. (2000). Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on visualization and computer graphics*, 6(1):24–43.
- [79] Hlawatsch, M., Burch, M., and Weiskopf, D. (2014). Visual adjacency lists for dynamic graphs. *IEEE Trans. Vis. and Comput. Graph.*, 20(11):1590–1603.
- [80] Holliman, N. S., Coltekin, A., Fernstad, S. J., McLaughlin, L., Simpson, M. D., and Woods, A. J. (2024). Entropy ordered shapes as bivariate glyphs. *Electronic Imaging*, 36(11).
- [81] Holten, D. (2006). Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Trans. Vis. Comput. Graphics*, 12(5):741–748.

- [82] Holten, D., Isenberg, P., Van Wijk, J. J., and Fekete, J.-D. (2011). An extended evaluation of the readability of tapered, animated, and textured directed-edge representations in node-link graphs. In *Proceedings of the 2011 IEEE Pacific Visualization Symposium*, pages 195–202. IEEE.
- [83] Holten, D. and Van Wijk, J. J. (2009). Force-directed edge bundling for graph visualization. *Computer graphics forum*, 28(3):983–990.
- [84] Hornbæk, K., Bederson, B. B., and Plaisant, C. (2002). Navigation patterns and usability of zoomable user interfaces with and without an overview. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 9(4):362–389.
- [85] House, D., Bair, A., and Ware, C. (2005). On the optimization of visualizations of complex phenomena. In *IEEE Visualization 2005*, pages 87–94. IEEE.
- [86] Huang, W., Eades, P., and Hong, S.-H. (2009). Measuring effectiveness of graph visualizations: A cognitive load perspective. *Information Visualization*, 8(3):139–152.
- [87] Hullman, J. (2016a). Why evaluating uncertainty visualization is error prone. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, BELIV '16, page 143–151.
- [88] Hullman, J. (2016b). Why evaluating uncertainty visualization is error prone. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pages 143–151.
- [89] Hunter, G. J. and Goodchild, M. (1993). Managing uncertainty in spatial databases: Putting theory into practice. *URISA Journal*, 5(2):55–62.
- [90] Jena, A., Engelke, U., Dwyer, T., Raiamanickam, V., and Paris, C. (2020a). Uncertainty visualisation: An interactive visual survey. *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pages 201–205.
- [91] Jena, A., Engelke, U., Dwyer, T., Raiamanickam, V., and Paris, C. (2020b). Uncertainty visualisation: An interactive visual survey. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*, pages 201–205. IEEE.
- [92] Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.
- [93] Jia, Y., Hoberock, J., Garland, M., and Hart, J. (2008). On the visualization of social and other scale-free networks. *IEEE transactions on visualization and computer graphics*, 14(6):1285–1292.
- [94] Johnson, B. and Shneiderman, B. (1991). Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In *Proceedings of the IEEE Conference on Visualization (Vis '91)*, pages 284–291.
- [95] Joslyn, S. and Savelli, S. (2010). Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteorological Applications*, 17(2):180–195.
- [96] Joslyn, S. L. and LeClerc, J. E. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of experimental psychology: applied*, 18(1):126.
- [97] Kamal, A., Dhakal, P., Javaid, A. Y., Devabhaktuni, V. K., Kaur, D., Zaiantz, J., and Marinier, R. (2021). Recent advances and challenges in uncertainty visualization: a survey. *Journal of visualization*, 24(5):861–890.

- [98] Karim, R. M., Kwon, O.-H., Park, C., and Lee, K. (2019). A study of colormaps in network visualization. *Applied Sciences*, 9(20):4228.
- [99] Karloff, H. and Shirley, K. E. (2013). Maximum entropy summary trees. *Computer graphics forum*, 32(3pt1):71–80.
- [100] Kassiano, V., Gounaris, A., Papadopoulos, A. N., and Tsihclas, K. (2016). Mining uncertain graphs: An overview. In *International Workshop on Algorithmic Aspects of Cloud Computing*, page 87–116.
- [101] Keim, D. A., Mansmann, F., Schneidewind, J., Thomas, J., and Ziegler, H. (2008). *Visual analytics: Scope and challenges*. Springer.
- [102] Keim, D. A., Mansmann, F., Schneidewind, J., and Ziegler, H. (2006). Challenges in visual data analysis. In *Tenth International Conference on Information Visualisation (IV'06)*, pages 9–16. IEEE.
- [103] Keller, R., Eckert, C. M., and Clarkson, P. J. (2006). Matrices or node-link diagrams: Which visual representation is better for visualising connectivity models? *Information Visualization*, 5(1):62–76.
- [104] Kerren, A., Purchase, H. C., and Ward, M. O., editors (2014). *Introduction to Multivariate Network Visualization*, volume 8380 of *Lecture Notes in Computer Science*. Springer.
- [105] Khan, K.-U., Nawaz, W., and Lee, Y.-K. (2015). Set-based approximate approach for lossless graph summarization. *Computing*, 97(12):1185–1207.
- [106] Khan, K. U., Nawaz, W., and Lee, Y.-K. (2017). Set-based unified approach for summarization of a multi-attributed graph. *World Wide Web*, 20(3):543–570.
- [107] Ko, S., Afzal, S., Walton, S., Yang, Y., Chae, J., Malik, A., Jang, Y., Chen, M., and Ebert, D. (2014). Analyzing high-dimensional multivariate network links with integrated anomaly detection, highlighting and exploration. In *Proceedings of the 2014 IEEE conference on visual analytics science and technology (VAST)*, pages 83–92. IEEE.
- [108] Kobourov, S. G., Mchedlidze, T., and Vonessen, L. (2015). Gestalt principles in graph drawing. In *Graph Drawing and Network Visualization: 23rd International Symposium, GD 2015, Los Angeles, CA, USA, September 24-26, 2015, Revised Selected Papers 23*, pages 558–560. Springer.
- [109] Koch, K., McLean, J., Segev, R., Freed, M. A., Berry II, M. J., Balasubramanian, V., and Sterling, P. (2006). How much the eye tells the brain. *Current biology*, 16(14):1428–1434.
- [110] Koffka, K. (1935). *Principles of Gestalt Psychology*. New York: Harcourt, Brace.
- [111] Koutra, D., Kang, U., Vreeken, J., and Faloutsos, C. (2015). Summarizing and understanding large graphs. *Stat. Anal. Data Min.*, 8(3):183–202.
- [112] Laramée, R. S. and Kosara, R. (2006). Challenges and unsolved problems. In *Human-Centered Visualization Environments*, page 231–254.
- [113] Lee, B., Plaisant, C., Parr, C. S., Fekete, J.-D., and Henry, N. (2006). Task taxonomy for graph visualization. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, BELIV '06, page 1–5, New York, NY, USA. Association for Computing Machinery.
- [114] LeFevre, K. and Terzi, E. (2010). Grass: Graph structure summarization. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 454–465. SIAM.

- [115] Li, E. Y., Liao, C. H., and Yen, H. R. (2013). Co-authorship networks and research impact: A social capital perspective. *Research Policy*, 42(9):1515–1530.
- [116] Lin, S., Fortuna, J., Kulkarni, C., Stone, M., and Heer, J. (2013). Selecting semantically-resonant colors for data visualization. *Computer Graphics Forum*, 32(3pt4):401–410.
- [117] Liu, S., Cui, W., Wu, Y., and Liu, M. (2014). A survey on information visualization: recent advances and challenges. *The Visual Computer*, 30(12):1373–1393.
- [118] Liu, W., Kan, A., Chan, J., Bailey, J., Leckie, C., Pei, J., and Kotagiri, R. (2012). On compressing weighted time-evolving graphs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2319–2322.
- [119] Liu, Y., Safavi, T., Dighe, A., and Koutra, D. (2018). Graph summarization methods and applications: A survey. *ACM computing surveys (CSUR)*, 51(3):1–34.
- [120] Lohse, G. L. (1997). The role of working memory on graphical information processing. *Behaviour & Information Technology*, 16(6):297–308.
- [121] MacEachren, A. M. (1992). Visualizing uncertain information. *Cartographic perspectives*, (13):10–19.
- [122] MacEachren, A. M., Roth, R. E., O’Brien, J., Li, B., Swingley, D., and Gahegan, M. (2012). Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2496–2505.
- [123] Magnello, M. E. (2009). Karl pearson and the establishment of mathematical statistics. *International Statistical Review*, 77(1):3–29.
- [124] Maguire, E., Rocca-Serra, P., Sansone, S.-A., Davies, J., and Chen, M. (2012). Taxonomy-based glyph design—with a case study on visualizing workflows of biological experiments. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2603–2612.
- [125] Martin, S., Brown, W. M., and Wylie, B. N. (2007). Drl: distributed recursive (graph) layout. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).
- [126] Mazza, R. (2009). *Introduction to Information Visualization*. Springer London, Limited, London.
- [127] McDonald, J. H. (2022). Handbook of biological statistics. <https://www.biostathandbook.com/exactgof.html>. Accessed Online: (01/02/2024).
- [128] McGee, F. and Dingliana, J. (2012). An empirical study on the impact of edge bundling on user comprehension of graphs. In *Proceedings of the International Working Conference on Advanced Visual Interfaces, AVI ’12*, page 620–627, New York, NY, USA. Association for Computing Machinery.
- [129] Meza, J., Xu, T., Veeraraghavan, K., and Mutlu, O. (2018). A large scale study of data center network reliability. In *Proceedings of the Internet Measurement Conference 2018*, pages 393–407.
- [130] Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97.
- [131] Mowshowitz, A. and Dehmer, M. (2012). Entropy and the complexity of graphs revisited. *Entropy*, 14(3):559–570.
- [132] Munzner, T. (1997). H3: Laying out large directed graphs in 3d hyperbolic space. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis ’97)*, pages 2–10. IEEE.

- [133] Munzner, T. (2014). *Visualization Analysis and Design*. AK Peters Visualization Series. CRC Press.
- [134] Napierala, M. A. (2012). What is the bonferroni correction? *Aaos Now*, pages 40–41.
- [135] Nobre, C., Meyer, M., Streit, M., and Lex, A. (2019). The state of the art in visualizing multivariate networks. *Computer Graphics Forum*, 38(3):807–832.
- [136] Nobre, C., Wootton, D., Harrison, L., and Lex, A. (2020). Evaluating multivariate network visualization techniques using a validated design and crowdsourcing approach. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12.
- [137] Okoe, M., Jianu, R., and Kobourov, S. (2018). Node-link or adjacency matrices: Old question, new insights. *IEEE transactions on visualization and computer graphics*, 25(10):2940–2952.
- [138] Olston, C. and Mackinlay, J. D. (2002). Visualizing data with bounded uncertainty. In *Proceedings of IEEE Symposium on Information Visualization*, INFOVIS’02, pages 37–40.
- [139] Padilla, L., Kay, M., and Hullman, J. (2022). Uncertainty visualization. In Piegorsch, W., Levine, R. A., Zhang, H. H., and Lee, T. C. M., editors, *Computational Statistics in Data Science*, chapter 21, pages 405–421. Wiley, Oxford.
- [140] Palan, S. and Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.
- [141] Pang, A. T., Wittenbrink, C. M., and Lodha, S. K. (1997). Approaches to uncertainty visualization. *The Visual Computer*, 13:370–390.
- [142] Peirce, J., Gray, J., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., and Lindeløv, J. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, 51.
- [143] Peter Statistics (n.d.). Analysing a binary variable, effect size: Cohen’s g . <https://peterstatistics.com/CrashCourse/2-SingleVar/Binary/Binary-2b-EffectSize.html>. Accessed: 2023-03-29.
- [144] Pienta, R., Abello, J., Kahng, M., and Chau, D. H. (2015). Scalable graph exploration and visualization: Sensemaking challenges and opportunities. In *Proceedings of the 2015 International conference on Big Data and smart computing (BIGCOMP)*, pages 271–278. IEEE.
- [145] Pohl, M., Schmitt, M., and Diehl, S. (2009). Comparing the readability of graph layouts using eyetracking and task-oriented analysis. In *Proceedings of the Fifth Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging*, Computational Aesthetics’09, page 49–56, Goslar, DEU. Eurographics Association.
- [146] Purchase, H. (1997). Which aesthetic has the greatest effect on human understanding? In *Graph Drawing*, volume 97, pages 248–261. Springer.
- [147] Purchase, H. C. (1998). Performance of layout algorithms: Comprehension, not computation. *Journal of Visual Languages & Computing*, 9(6):647–657.
- [148] Purchase, H. C., Cohen, R. F., and James, M. (1995). Validating graph drawing aesthetics. In *Proceedings of the Graph Drawing Symposium*, pages 435–446. Springer.
- [149] Quinlan, P. T. and Humphreys, G. W. (1987). Visual search for targets defined by combinations of color, shape, and size: An examination of the task constraints on feature and conjunction searches. *Perception & Psychophysics*, 41:455–472.

- [150] Ramos, M. H., Van Andel, S. J., and Pappenberger, F. (2013). Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences*, 17(6):2219–2232.
- [151] Riondato, M., García-Soriano, D., and Bonchi, F. (2014). Graph summarization with quality guarantees. In *Proceedings of the 2014 IEEE International Conference on Data Mining, ICDM '14*, page 947–952. IEEE Computer Society.
- [152] Roberts, J. C., Yang, J., Kohlbacher, O., Ward, M. O., and Zhou, M. X. (2014). Novel visual metaphors for multivariate networks. In Kerren, A., Purchase, H. C., and Ward, M. O., editors, *Multivariate Network Visualization*, chapter 7, pages 127–150. Springer International Publishing.
- [153] Rogowitz, B. E. and Kalvin, A. D. (2001). The “which blair project” : A quick visual method for evaluating perceptual color maps. In *Proceedings Visualization, 2001. VIS'01.*, pages 183–556. IEEE.
- [154] Romat, H., Appert, C., Bach, B., Henry-Riche, N., and Pietriga, E. (2018). Animated edge textures in node-link diagrams: A design space and initial evaluation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18*, page 1–13, New York, NY, USA. Association for Computing Machinery.
- [155] Ropinski, T., Oeltze-Jafra, S., and Preim, B. (2011). Survey of glyph-based visualization techniques for spatial multivariate medical data. *Computers & Graphics*, 35:392–401.
- [156] Rosenholtz, R., Li, Y., Mansfield, J., and Jin, Z. (2005). Feature congestion: A measure of display clutter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '05*, page 761–770, New York, NY, USA. Association for Computing Machinery.
- [157] Rosenthal, R., Cooper, H., and Hedges, L. V. (1994). Parametric measures of effect size. In *The handbook of research synthesis*, pages 231–244. Russell Sage Foundation.
- [158] Roulston, M. S., Bolton, G. E., Kleit, A. N., and Sears-Collins, A. L. (2006). A laboratory study of the benefits of including uncertainty information in weather forecasts. *Weather and Forecasting*, 21(1):116–122.
- [159] Saraiya, P., Lee, P., and North, C. (2005). Visualization of graphs with associated timeseries data. In *Proceedings of the 2005 IEEE Symposium on Information Visualization (INFOVIS'05)*, pages 225–232. IEEE.
- [160] Sayama, H. (2015). *Introduction to the modeling and analysis of complex systems*. Open SUNY Textbooks.
- [161] Schöffel, S., Schwank, J., and Ebert, A. (2016). A user study on multivariate edge visualizations for graph-based visual analysis tasks. In *Proceedings of the 2016 20th International Conference Information Visualisation (IV)*, pages 165–170. IEEE.
- [162] Schulz, H.-J., Nocke, T., Heitzler, M., and Schumann, H. (2013). A design space of visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2366–2375.
- [163] Schwank, J., Schöffel, S., Stärz, J., and Ebert, A. (2016). Visualizing uncertainty of edge attributes in node-link diagrams. *2016 20th International Conference Information Visualisation (IV)*, pages 45–50.
- [164] Seidler, P., Haider, J., Kodagoda, N., Wong, B. W., Pohl, M., and Adderley, R. (2016). Design for intelligence analysis of complex systems: evolution of criminal networks. In *Proceedings of the European Intelligence and Security Informatics Conference*, pages 140–143. IEEE.

- [165] Shah, N., Koutra, D., Jin, L., Zou, T., Gallagher, B., and Faloutsos, C. (2017). On summarizing large-scale dynamic graphs. *IEEE Data Eng. Bull.*, 40(3):75–88.
- [166] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.
- [167] Shannon, C. E. and Weaver, W. (1949). *The mathematical theory of communication*. University of Illinois Press.
- [168] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504.
- [169] Shen, Z., Ma, K.-L., and Eliassi-Rad, T. (2006). Visual analysis of large heterogeneous social networks by semantic and structural abstraction. *IEEE transactions on visualization and computer graphics*, 12(6):1427–1439.
- [170] Shi, L., Tong, H., Tang, J., and Lin, C. (2015). Vegas: Visual influence graph summarization on citation networks. *IEEE Transactions on Knowledge and Data Engineering*, 27(12):3417–3431.
- [171] Shiravi, H., Shiravi, A., and Ghorbani, A. A. (2012). A survey of visualization systems for network security. *IEEE Trans. Vis. Comput. Graphics*, 18(8):1313–1329.
- [172] Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE symposium on visual languages*, pages 336–343.
- [173] Shneiderman, B. and Aris, A. (2006). Network visualization by semantic substrates. *IEEE Trans. Vis. and Comput. Graph.*, 12(5):733–740.
- [174] Skiena, S. S. (2020). *Graph Traversal*, pages 197–242. Springer International Publishing.
- [175] Slingsby, A., Dykes, J., and Wood, J. (2009). Configuring hierarchical layouts to address research questions. *IEEE Trans. Vis. and Comput. Graph.*, 15(6):977–984.
- [176] Smart, S. and Szafir, D. A. (2019). Measuring the separability of shape, size, and color in scatterplots. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14.
- [177] Spence, R. (2001). *Information visualization*. ACM Press, Harlow.
- [178] Spiegelhalter, D. (2017). Risk and uncertainty communication. *Annual Review of Statistics and Its Application*, 4:31–60.
- [179] Stasko, J. and Zhang, E. (2000). Focus+ context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *Proceedings of the 2000 IEEE Symposium on Information Visualization*, pages 57–65. IEEE.
- [180] StatsTest (2022). G-test of goodness of fit. <https://www.statstest.com/g-test-of-goodness-of-fit/>. Accessed Online: (01/02/2024).
- [181] Sun, G.-D., Wu, Y.-C., Liang, R.-H., and Liu, S.-X. (2013). A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology*, 28(5):852–867.
- [182] Tamassia, R. (2013). *Handbook of graph drawing and visualization*. CRC press.
- [183] Tantipathananandh, C. and Berger-Wolf, T. Y. (2011). Finding communities in dynamic social networks. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, ICDM’11, pages 1236–1241. IEEE.

- [184] Telea, A. and Ersoy, O. (2010). Image-based edge bundles: Simplified visualization of large graphs. *Computer Graphics Forum*, 29(3):843–852.
- [185] Thomas, J. J. and Cook, K. A. (2006). A visual analytics agenda. *IEEE computer graphics and applications*, 26(1):10–13.
- [186] Tian, Y., Hankins, R. A., and Patel, J. M. (2008). Efficient aggregation for graph summarization. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 567–580.
- [187] Toivonen, H., Zhou, F., Hartikainen, A., and Hinkka, A. (2011). Compression of weighted graphs. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'11, page 965–973.
- [188] Treisman, A. (1985). Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing*, 31(2):156–177.
- [189] Tu, Y. and Shen, H.-W. (2007). Visualizing changes of hierarchical data using treemaps. *IEEE Trans. Vis. and Comput. Graph.*, 13(6):1286–1293.
- [190] Usher, M. J. (1984). *Information theory for information technologists*. Macmillan Press Ltd.
- [191] van den Elzen, S. and van Wijk, J. J. (2014). Multivariate network exploration and presentation: From detail to overview via selections and aggregations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2310–2319.
- [192] van Ham, F. and Perer, A. (2009). “search, show context, expand on demand”: Supporting large graph exploration with degree-of-interest. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):953–960.
- [193] Van Ham, F. and van Wijk, J. J. (2003). Beamtrees: Compact visualization of large hierarchies. *Information Visualization*, 2(1):31–39.
- [194] van Steen, M. (2010). *Graph Theory and Complex Networks: An Introduction*. Maarten van Steen.
- [195] Van Wijk, J. J. and Van de Wetering, H. (1999). Cushion treemaps: Visualization of hierarchical information. In *Proceedings of the 1999 IEEE Symposium on Information Visualization (InfoVis' 99)*, pages 73–78. IEEE.
- [196] von Landesberger, T., Bremm, S., and Wunderlich, M. (2017). Typology of uncertainty in static geolocated graphs for visualization. *IEEE Computer Graphics and Applications*, 37(5):18–27.
- [197] von Landesberger, T., Kuijper, A., Schreck, T., Kohlhammer, J., van Wijk, J., Fekete, J.-D., and Fellner, D. (2011). Visual analysis of large graphs: State-of-the-art and future research challenges. *Computer graphics forum*, 30(6):1719–1749.
- [198] Wallinger, M., Archambault, D., Auber, D., Nöllenburg, M., and Peltonen, J. (2022). Edge-path bundling: A less ambiguous edge bundling approach. *IEEE Trans. Vis. Comput. Graph.*, 28(1):313–323.
- [199] Wang, Y., Wang, Y., Zhang, H., Sun, Y., Fu, C.-W., Sedlmair, M., Chen, B., and Deussen, O. (2018). Structure-aware fisheye views for efficient large graph exploration. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):566–575.
- [200] Ward, M. O. (2002). A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, 1(3-4):194–210.

- [201] Ware, C. (2019). *Information Visualization : Perception for Design*. Morgan Kaufmann, fourth edition.
- [202] Ware, C. and Bobrow, R. (2005). Supporting visual queries on medium-sized node-link diagrams. *Information Visualization*, 4(1):49–58.
- [203] Ware, C., Purchase, H., Colpoys, L., and McGill, M. (2002). Cognitive measurements of graph aesthetics. *Information visualization*, 1(2):103–110.
- [204] Wattenberg, M. (2006). Visual exploration of multivariate graphs. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 811–819.
- [205] Weiskopf, D. (2022). Uncertainty visualization: Concepts, methods, and applications in biological data visualization. *Frontiers in Bioinformatics*, 2:793819.
- [206] Windhager, F., Salisu, S., and Mayr, E. (2019). Exhibiting uncertainty: Visualizing data quality indicators for cultural collections. *Informatics*, 6(3).
- [207] Wu, J., Zhu, F., Liu, X., and Yu, H. (2018). An information-theoretic framework for evaluating edge bundling visualization. *Entropy (Basel, Switzerland)*, 20(9):625.
- [208] Wu, Y., Cao, N., Archambault, D., Shen, Q., Qu, H., and Cui, W. (2016). Evaluation of graph sampling: A visualization perspective. *IEEE transactions on visualization and computer graphics*, 23(1):401–410.
- [209] Wybrow, M., Elmqvist, N., Fekete, J.-D., Von Landesberger, T., van Wijk, J. J., and Zimmer, B. (2014). Interaction in the visualization of multivariate networks. In Kerren, A., Purchase, H. C., and Ward, M. O., editors, *Multivariate Network Visualization*, chapter 6, pages 97–125. Springer International Publishing.
- [210] Xu, K., Rooney, C., Passmore, P., Ham, D.-H., and Nguyen, P. H. (2012). A user study on curved edges in graph visualization. *IEEE Trans. Vis. Comput. Graphics*, 18(12):2449–2456.
- [211] Yoghourdjian, V., Archambault, D., Diehl, S., Dwyer, T., Klein, K., Purchase, H. C., and Wu, H.-Y. (2018). Exploring the limits of complexity: A survey of empirical studies on graph visualisation. *Visual Informatics*, 2(4):264–282.
- [212] Zhang, N., Tian, Y., and Patel, J. M. (2010). Discovery-driven graph summarization. In *2010 IEEE 26th International Conference on Data Engineering (ICDE)*., pages 880–891. IEEE.
- [213] Zheng, B. and Sadlo, F. (2021). On the visualization of hierarchical multivariate data. In *Proceedings of the 2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*, pages 136–145. IEEE.
- [214] Zhou, Z., Shi, C., Shen, X., Cai, L., Wang, H., Liu, Y., Zhao, Y., and Chen, W. (2020). Context-aware sampling of large networks via graph representation learning. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1709–1719.
- [215] Zhu, M., Chen, W., Hu, Y., Hou, Y., Liu, L., and Zhang, K. (2020). Drgraph: An efficient graph layout algorithm for large-scale graphs by dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1666–1676.

Appendix A

Experiment 1

Stimuli of Vizent and the control group of visual encodings

Task 1 aims to identify which edge in the display graph has the lowest variability, regardless of its network traffic value.

Task 2 aims to identify which edge in the display graph has the highest network traffic, regardless of its variation.

Task 3 aims to identify which edge in the display graph has the highest network traffic value while also having the lowest variability in that traffic.

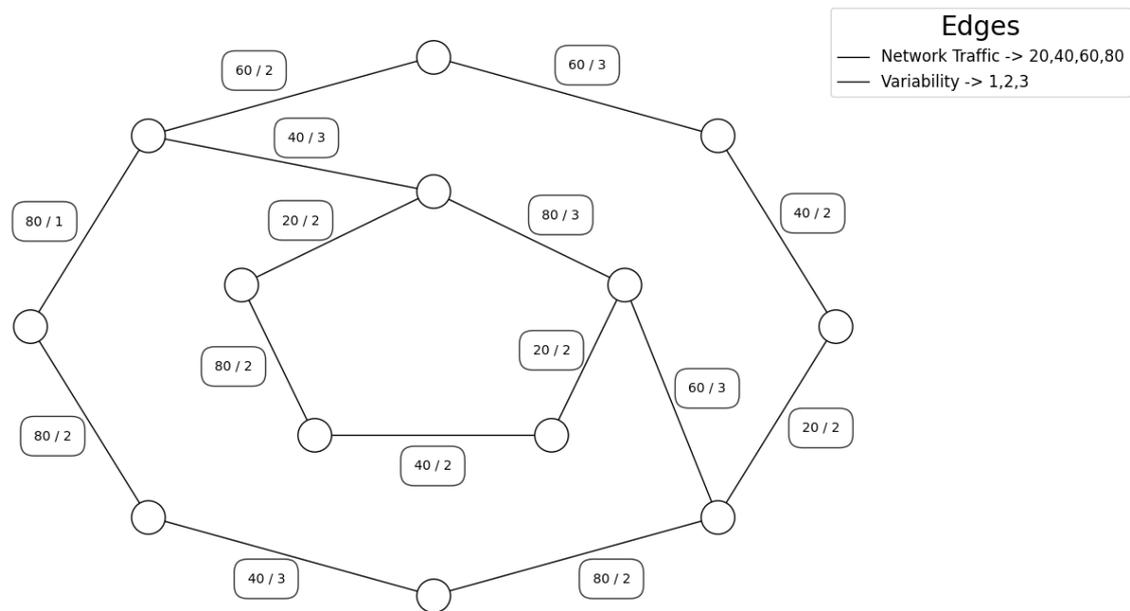


Figure A.1 A 15-edges graph used as stimuli of *Num* for Task 1

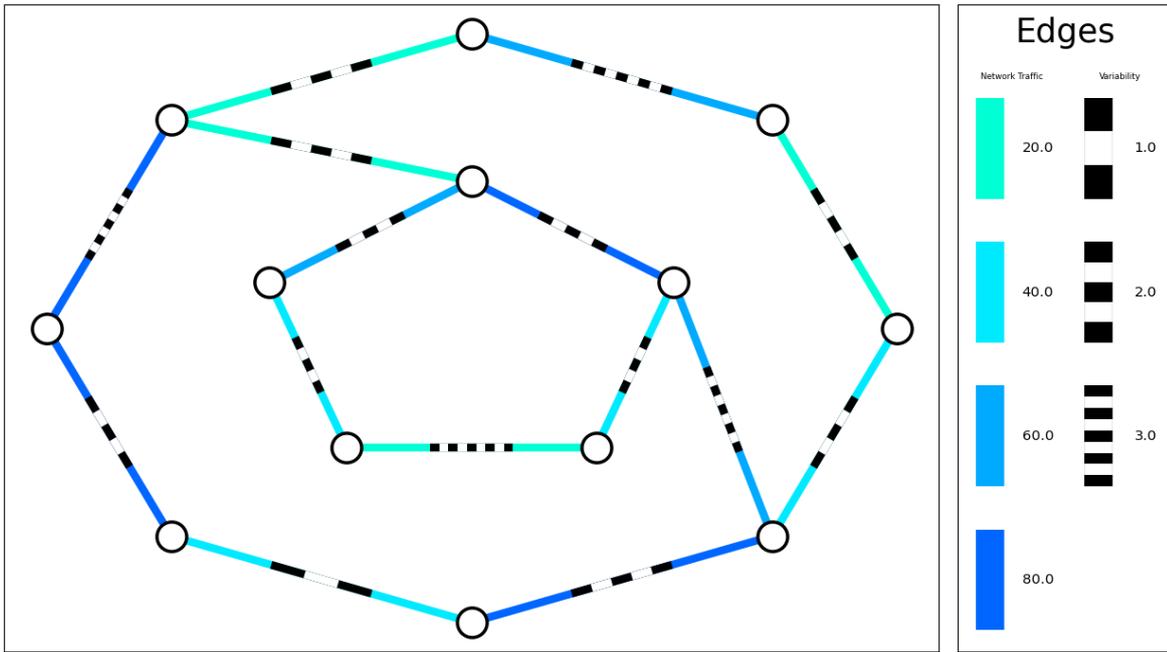


Figure A.2 A 15-edges graph used as stimuli of Vizent for Task 1

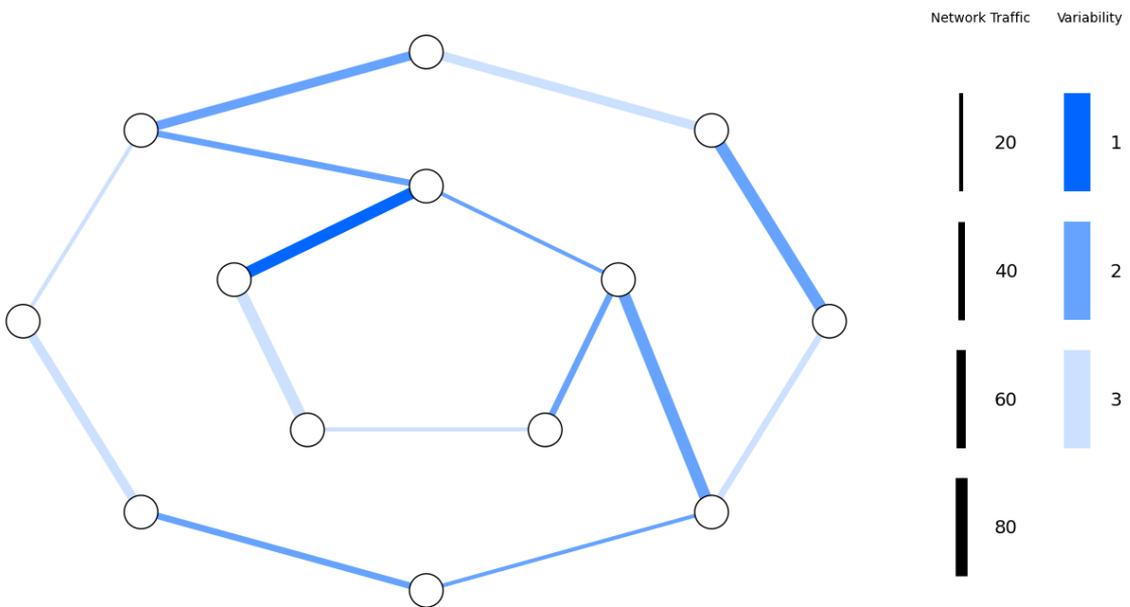


Figure A.3 A 15-edges graph used as stimuli of *Wid-Lig* for Task 1

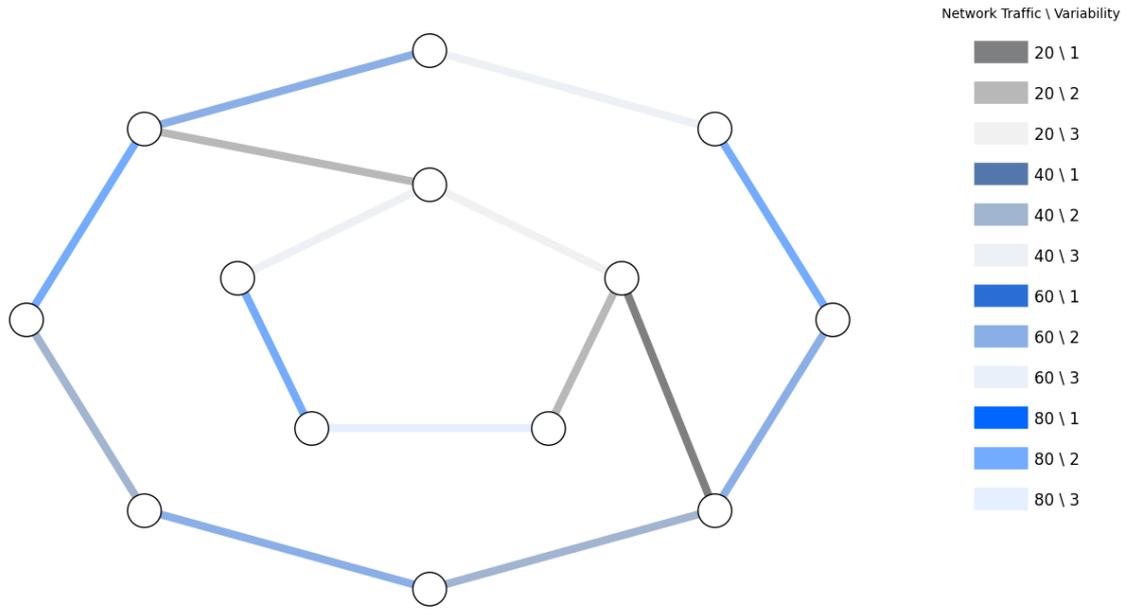


Figure A.4 A 15-edges graph used as stimuli of *Sat-Tra* for Task 1

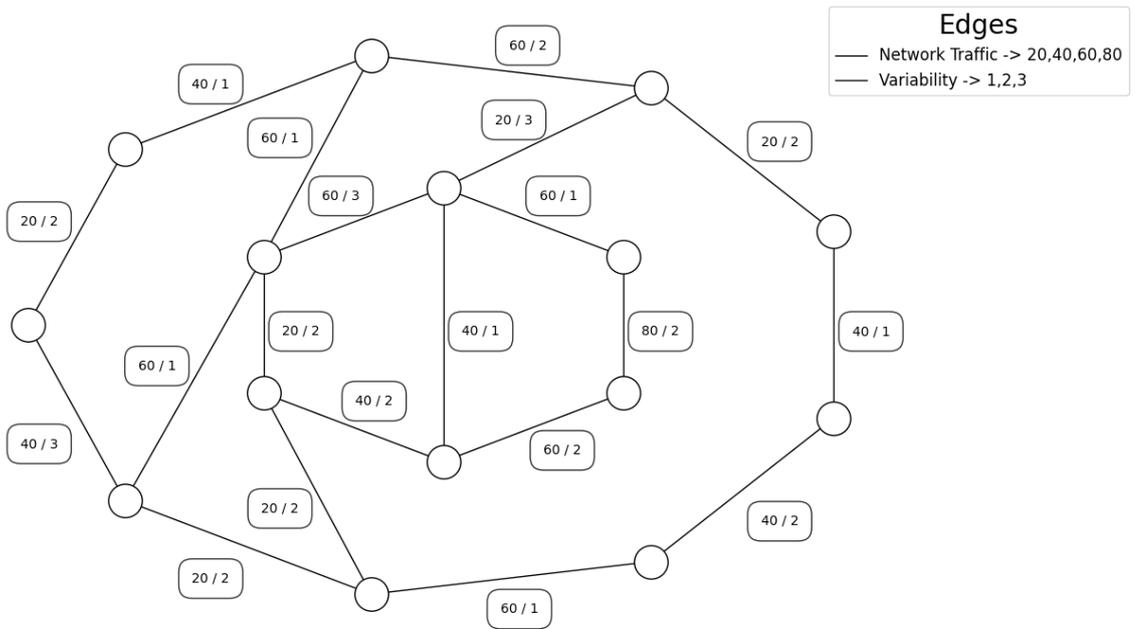


Figure A.5 A 20-edges graph used as stimuli of *Num* for Task 2

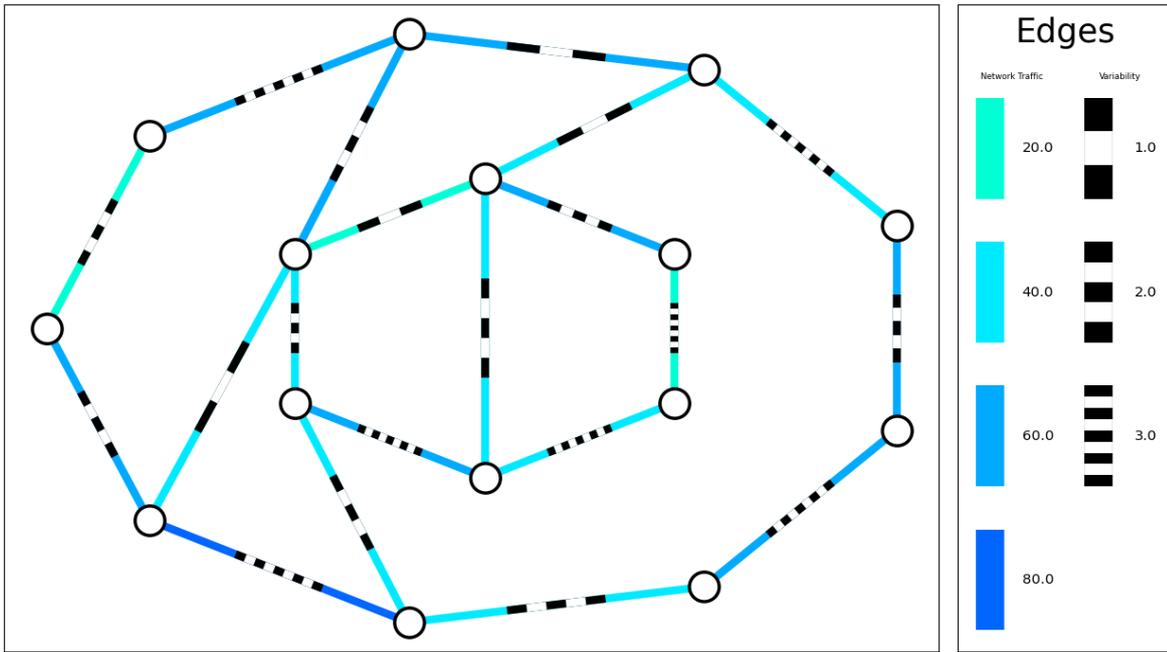


Figure A.6 A 20-edges graph used as stimuli of Vizent for Task 2

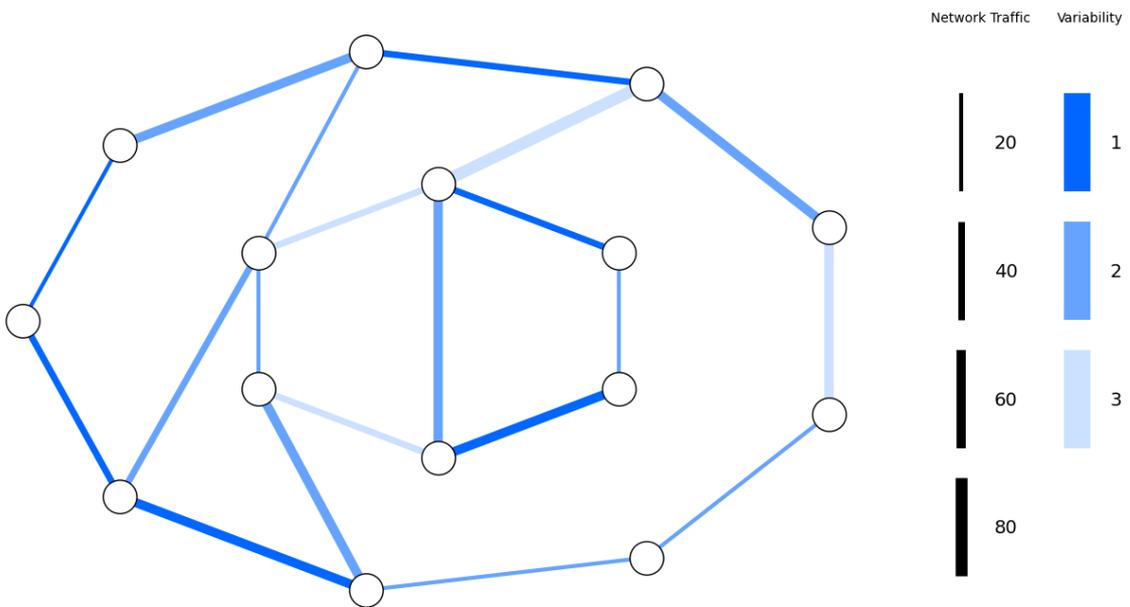


Figure A.7 A 20-edges graph used as stimuli of *Wid-Lig* for Task 2

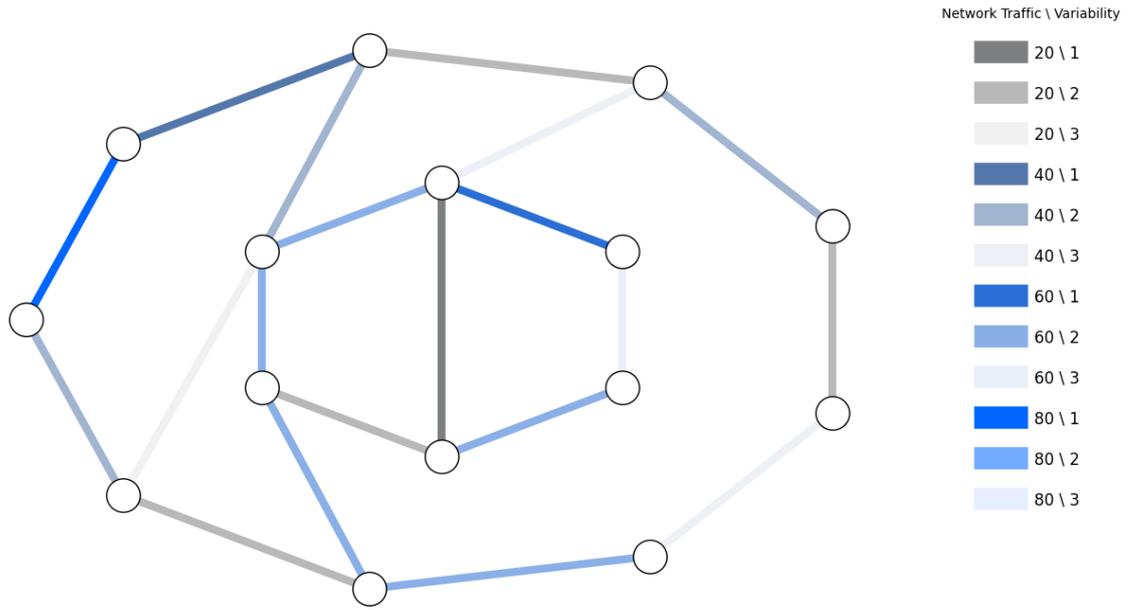


Figure A.8 A 20-edges graph used as stimuli of *Sat-Tra* for Task 2

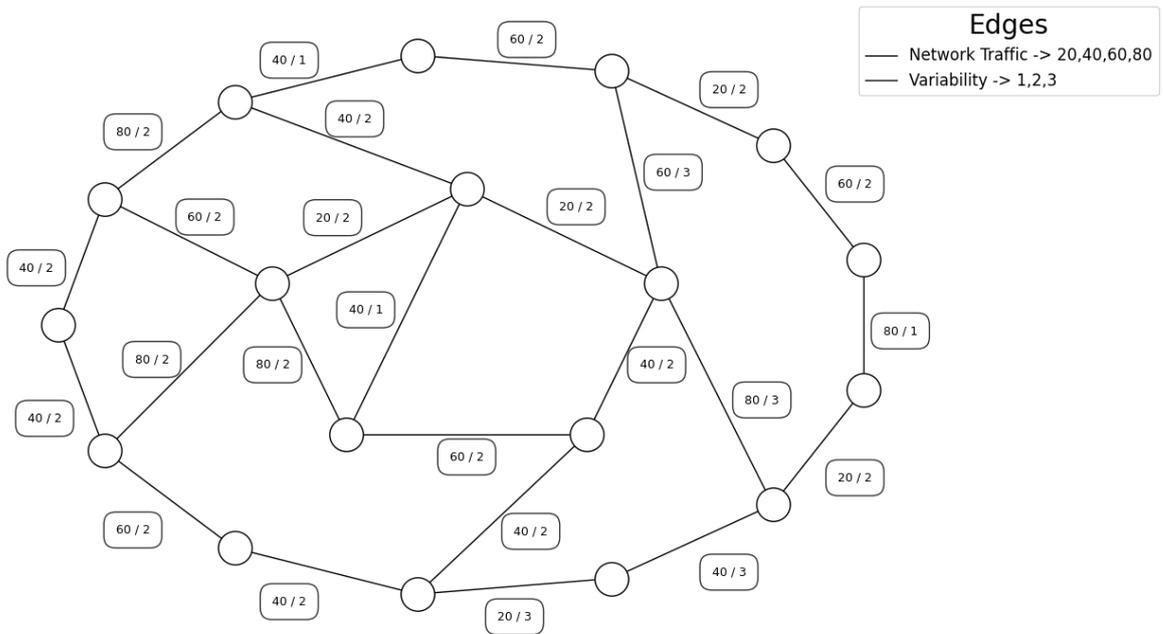


Figure A.9 A 25-edges graph used as stimuli of *Num* for Task 3

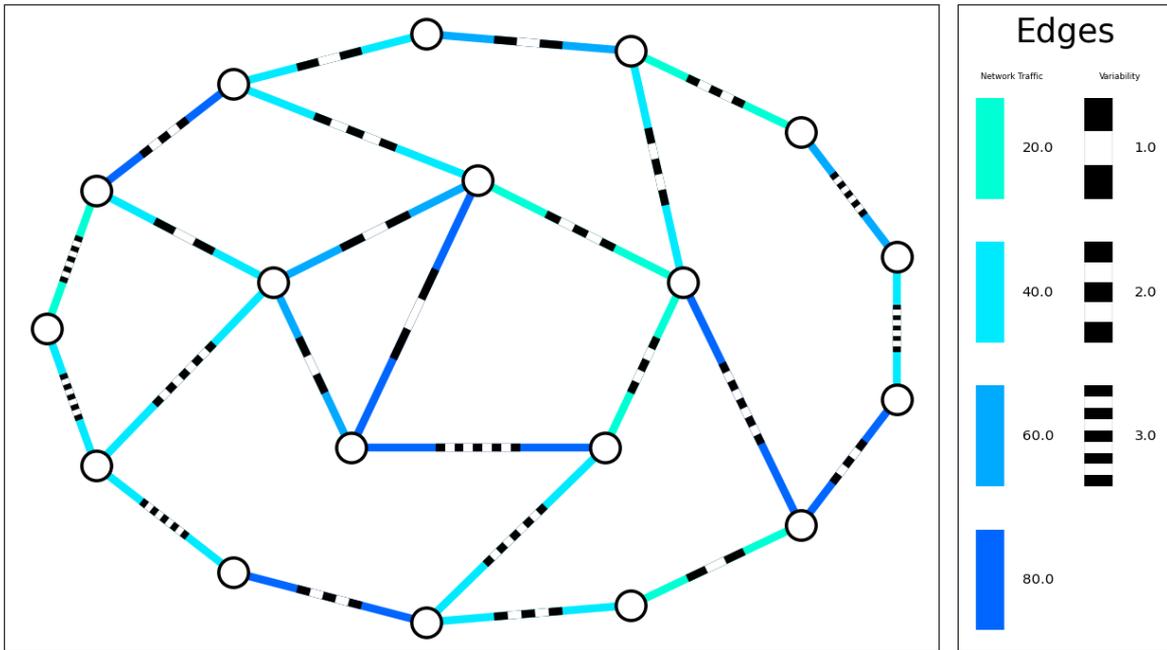


Figure A.10 A 25-edges graph used as stimuli of Vizent for Task 3

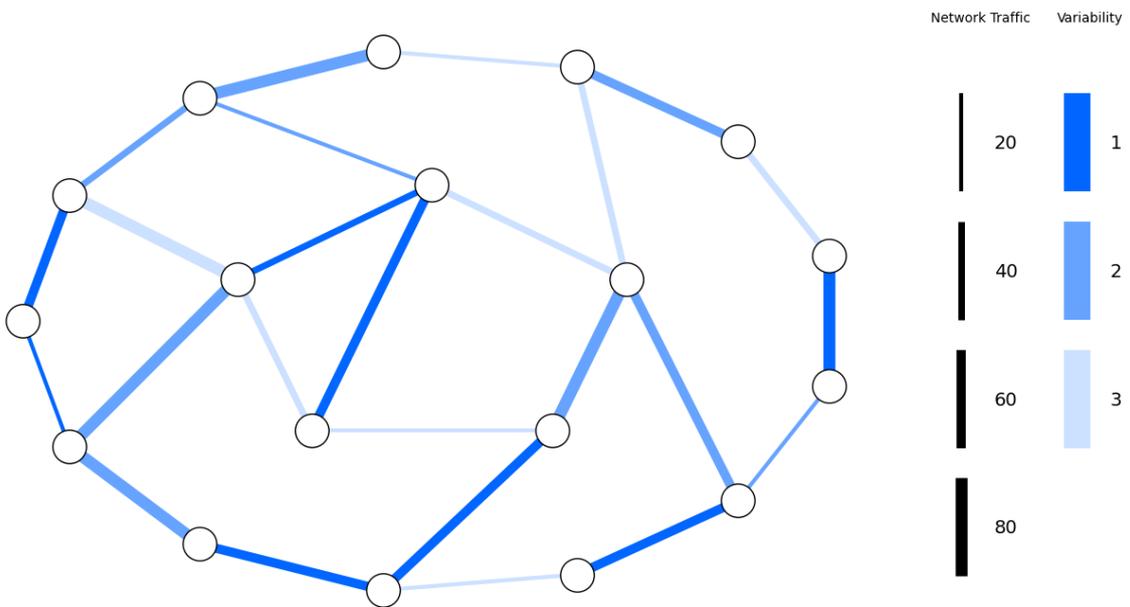


Figure A.11 A 25-edges graph used as stimuli of *Wid-Lig* for Task 3

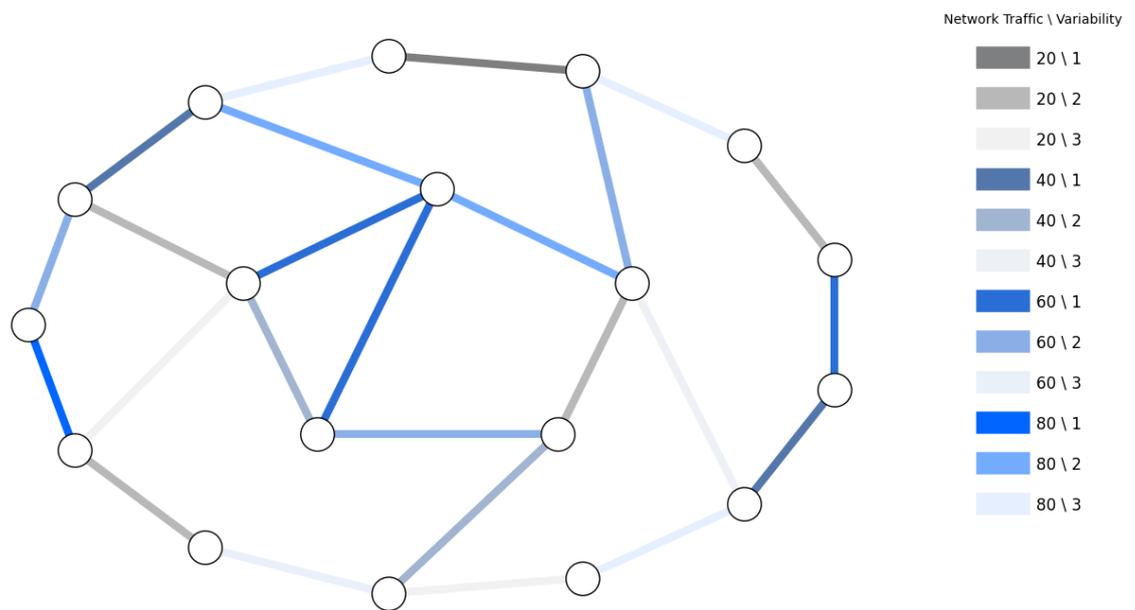


Figure A.12 A 25-edges graph used as stimuli of *Sat-Tra* for Task 3

Appendix B

Experiment-2

Stimuli of Vizent Graph and Numerical Values Visualisation

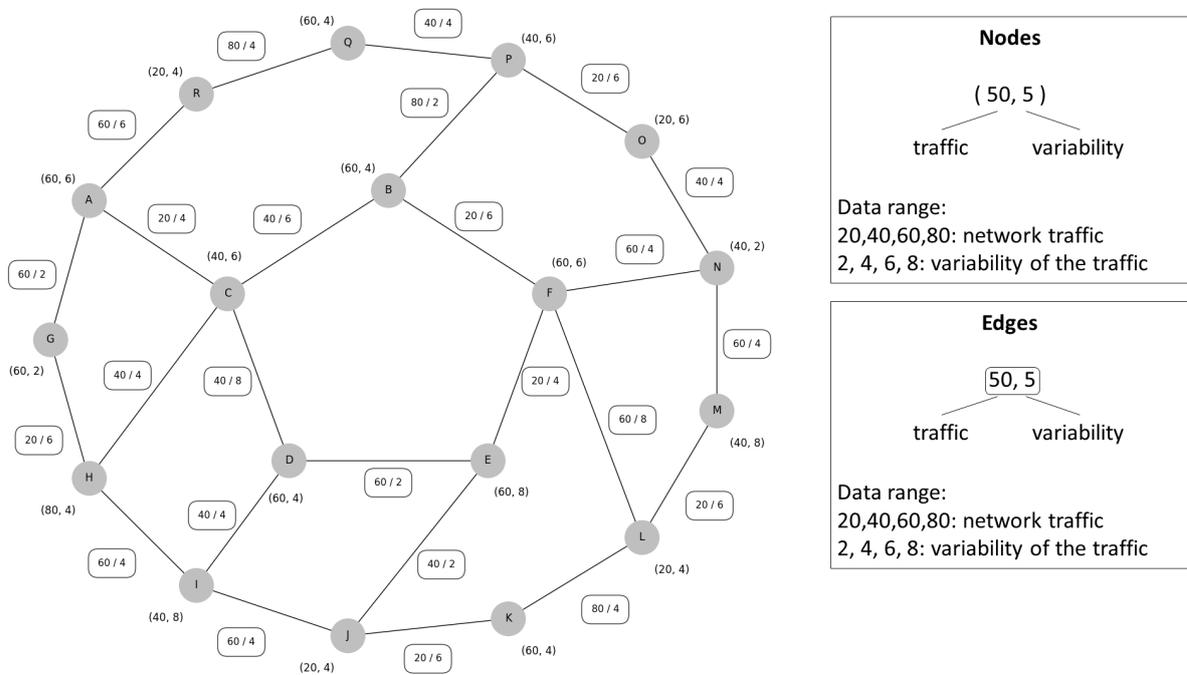


Figure B.1 An example of the Numerical values visualisation used as a stimulus in the experiment.

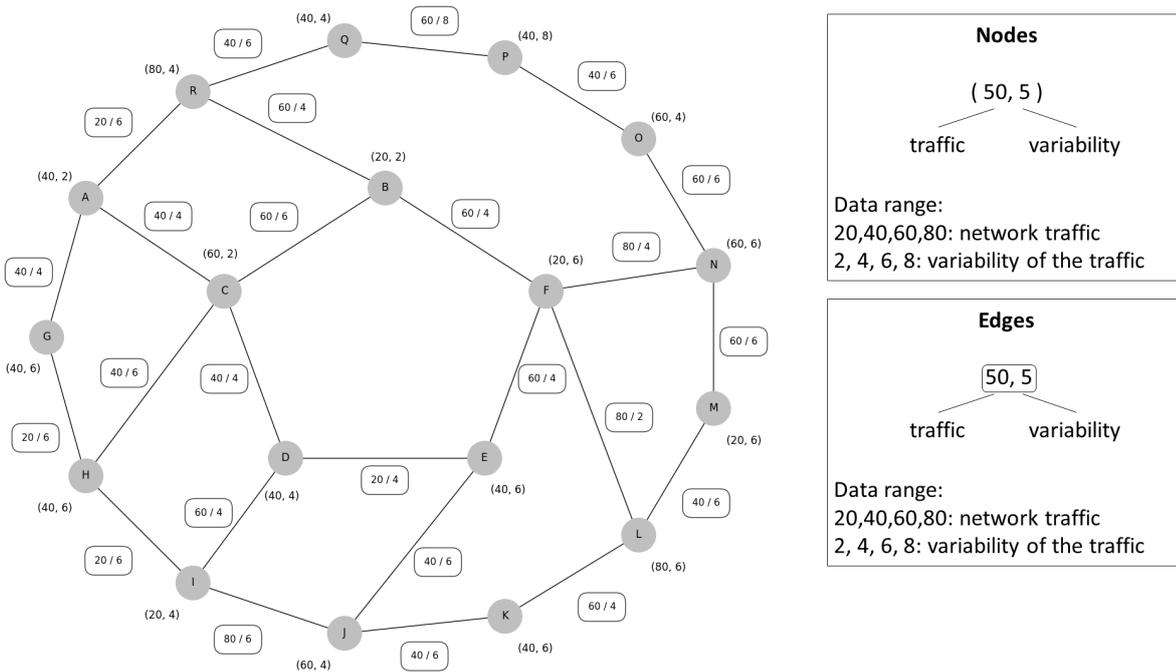


Figure B.2 An example of the Numerical values visualisation used as a stimulus in the experiment.

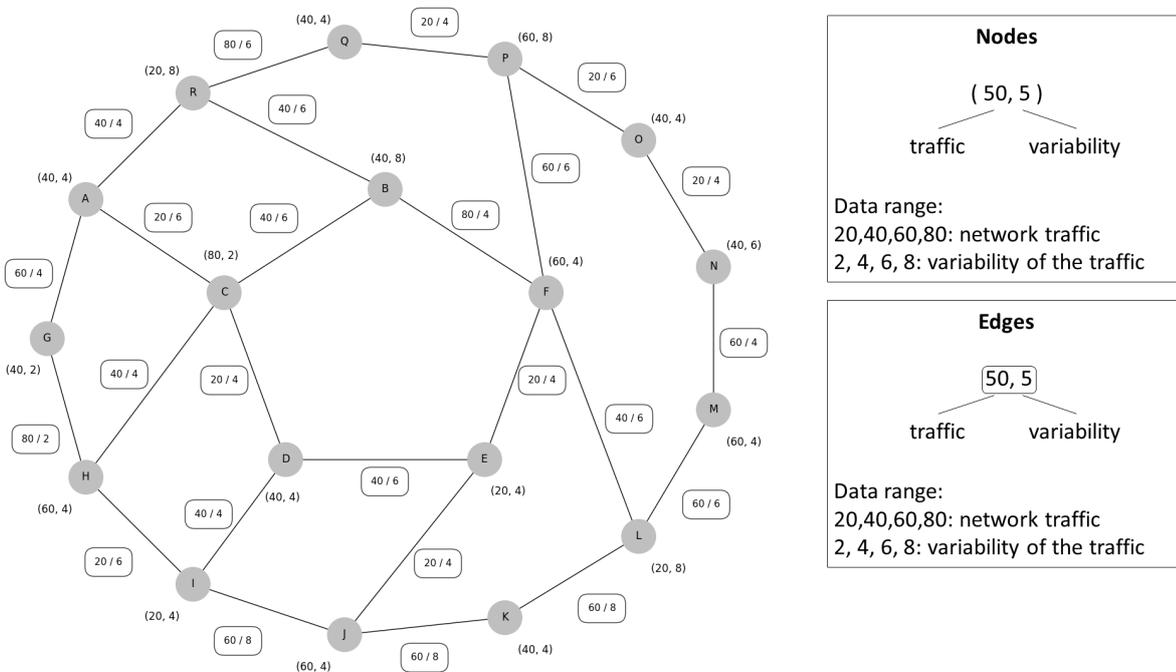


Figure B.3 An example of the Numerical values visualisation used as a stimulus in the experiment.

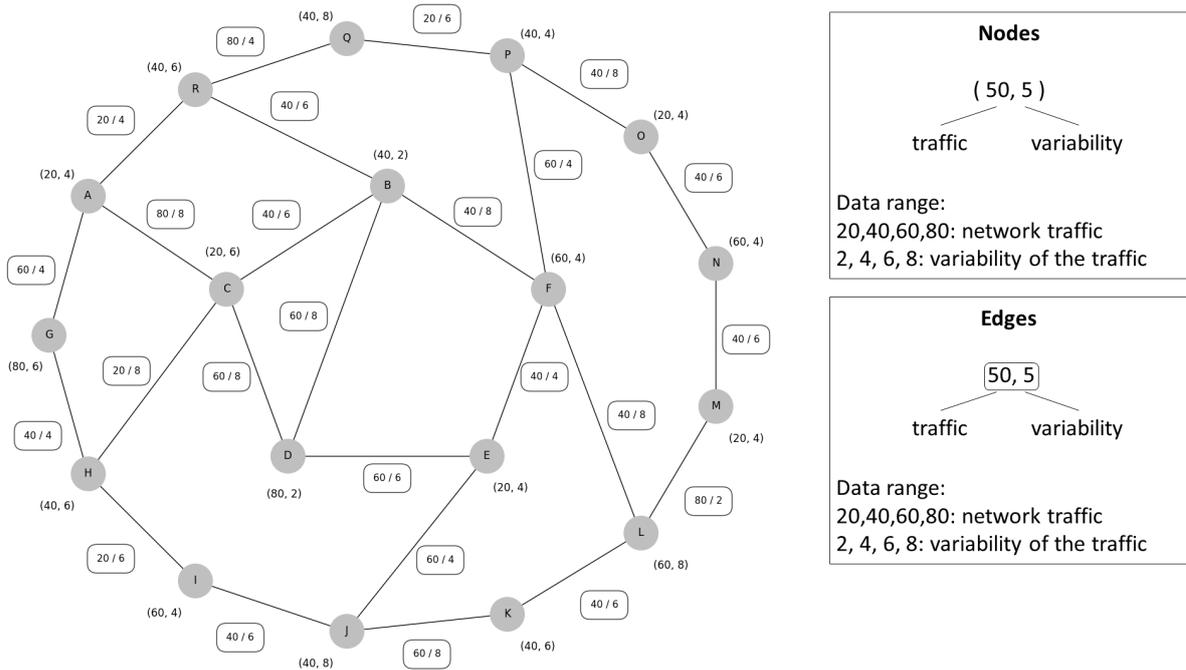


Figure B.4 An example of the Numerical values visualisation used as a stimulus in the experiment.

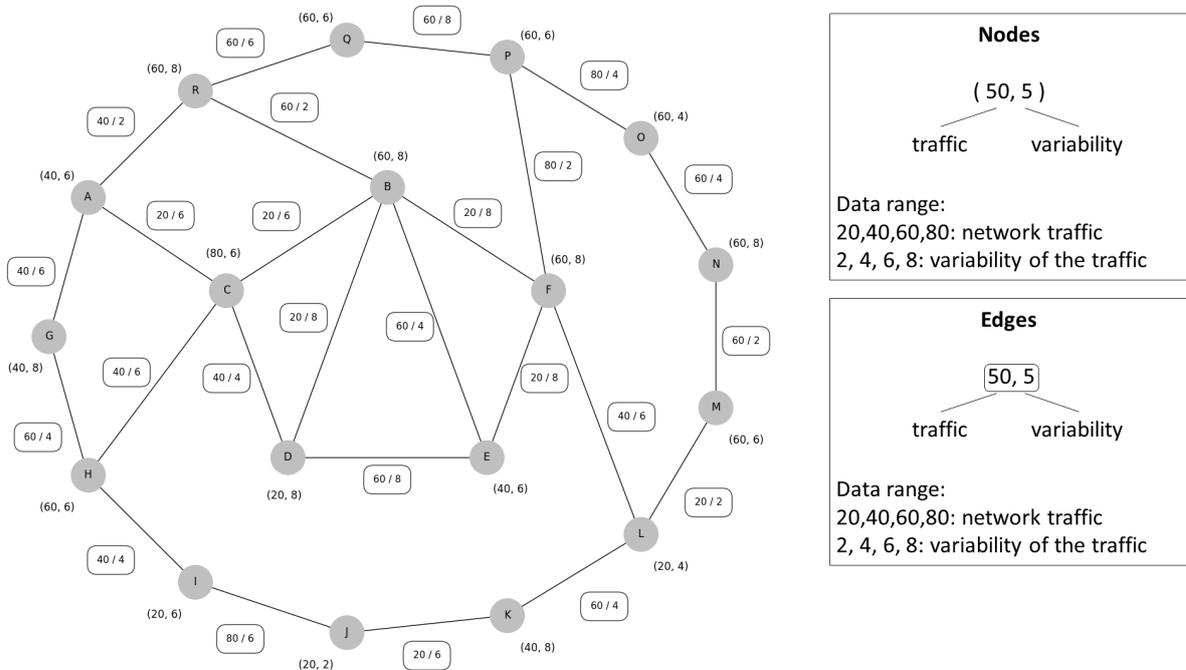


Figure B.5 An example of the Numerical values visualisation used as a stimulus in the experiment.

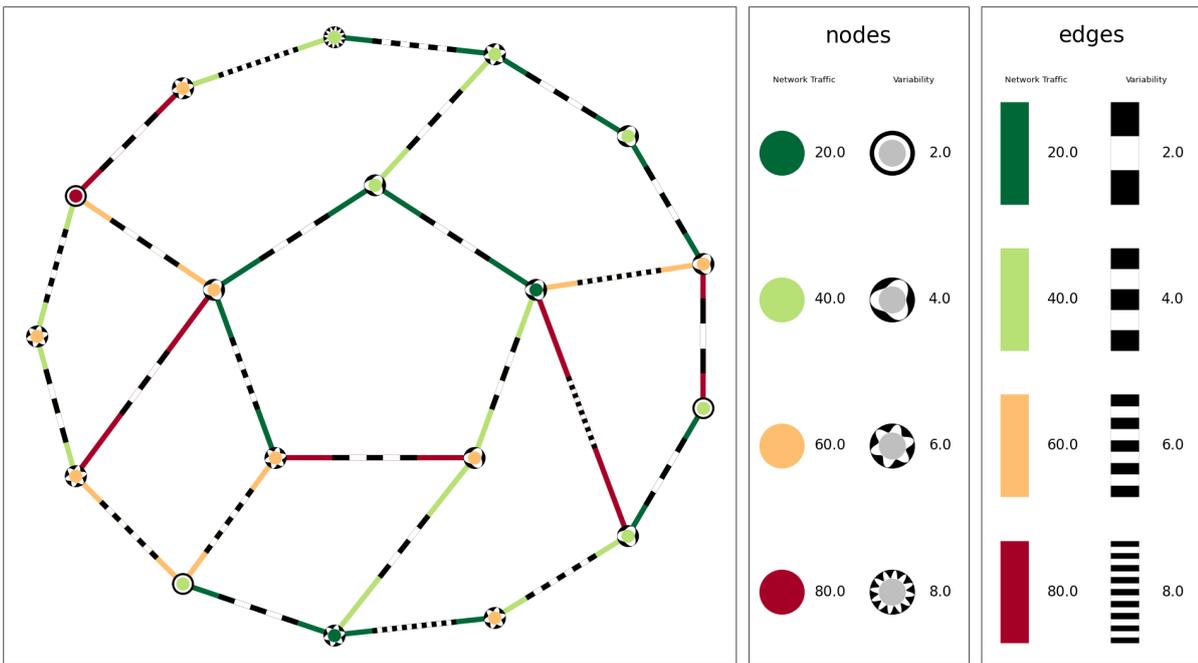


Figure B.6 An example of the Vizent graph used as a stimulus in the experiment.

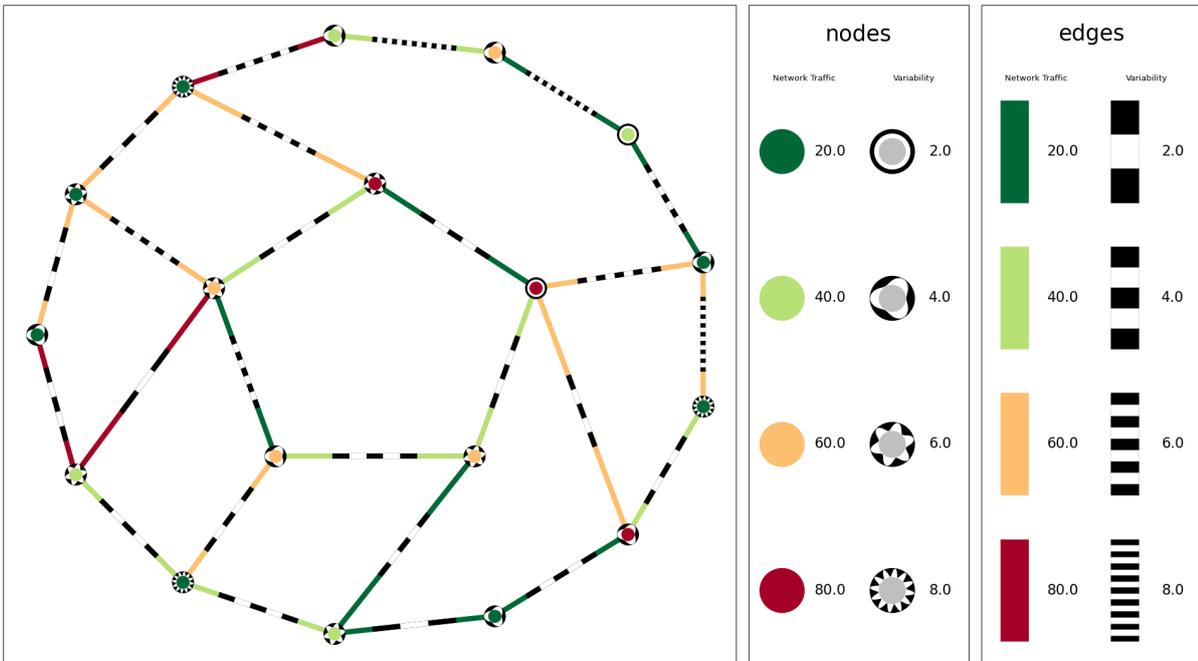


Figure B.7 An example of the Vizent graph used as a stimulus in the experiment.

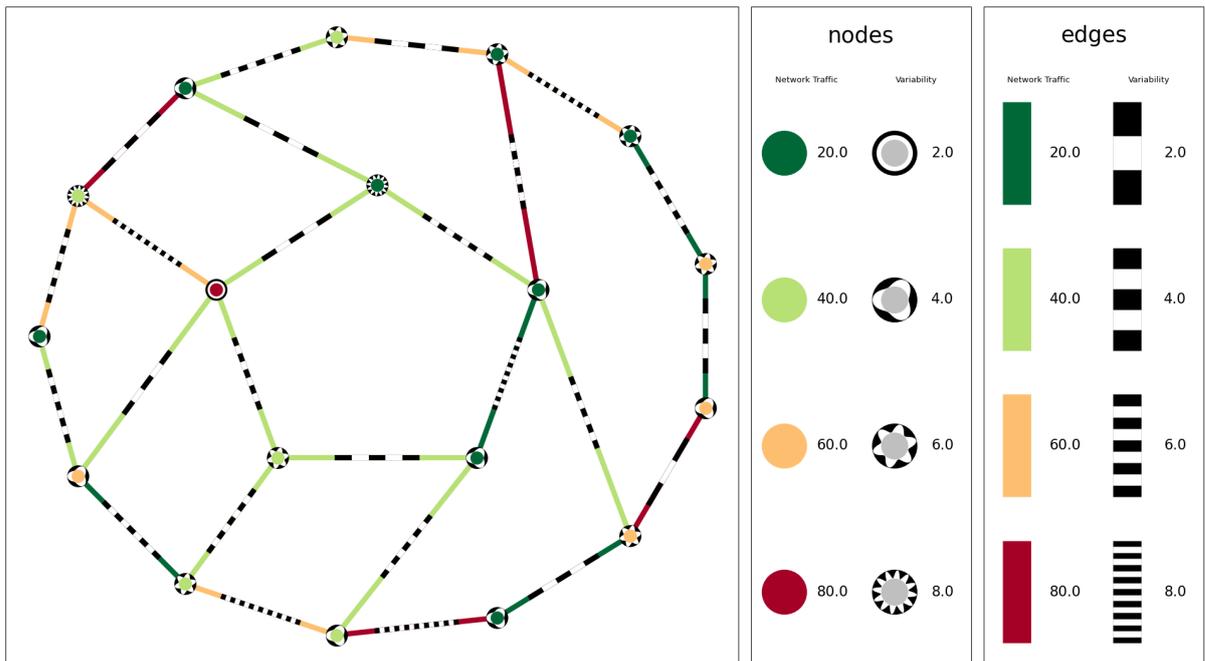


Figure B.8 An example of the Vizent graph used as a stimulus in the experiment.

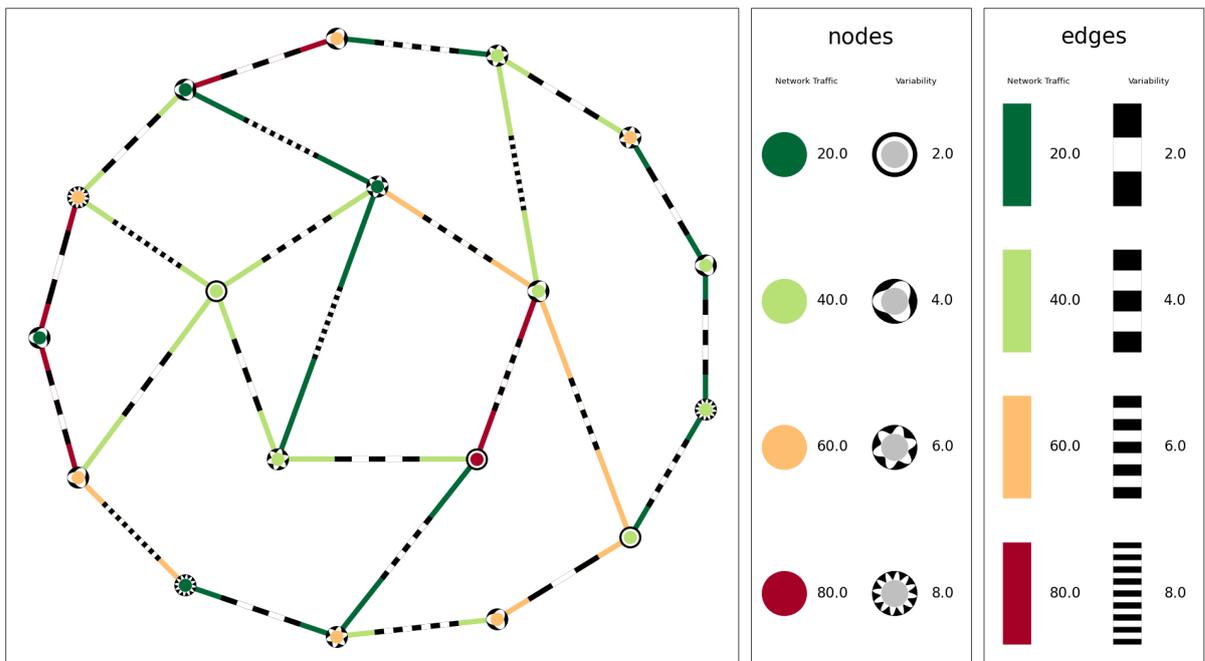


Figure B.9 An example of the Vizent graph used as a stimulus in the experiment.

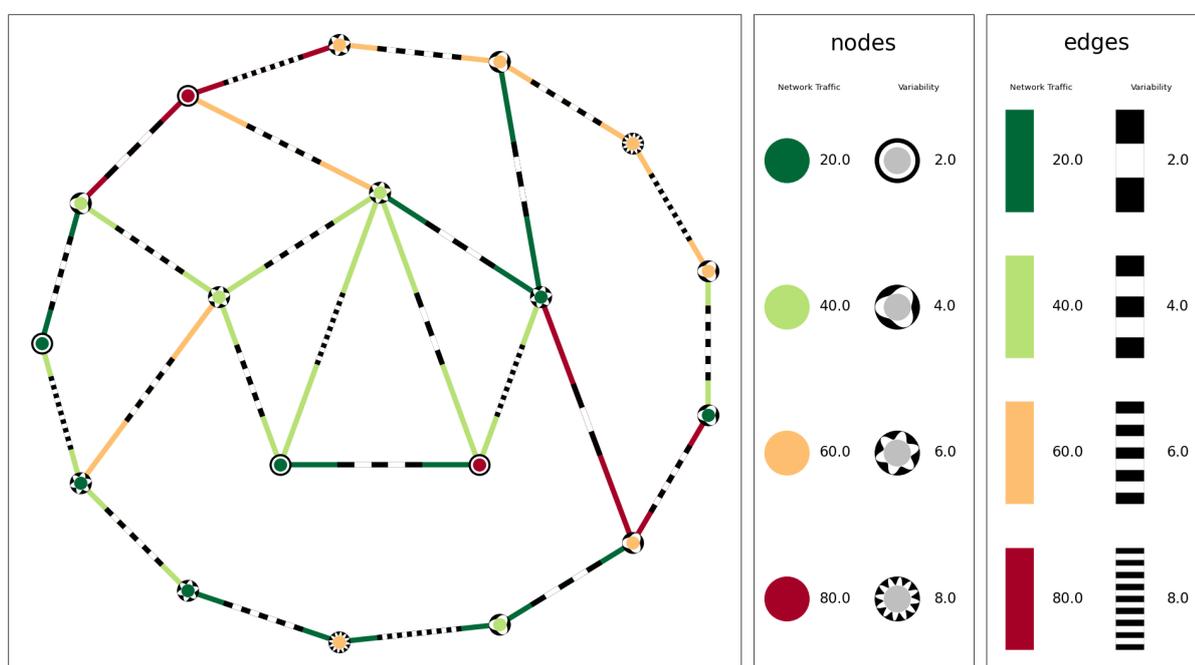


Figure B.10 An example of the Vizent graph used as a stimulus in the experiment.