



THESIS / THÈSE

MASTER IN BUSINESS ENGINEERING PROFESSIONAL FOCUS IN ANALYTICS & DIGITAL BUSINESS

How can graphical tools such as colors, shapes and positions of nodes or links influence readers' interpretation of graphs support to structural analysis ?

Tellier, Harold

Award date:
2021

Awarding institution:
University of Namur

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.



How can graphical tools such as colors, shapes and positions of nodes or links influence readers' interpretation of graphs: support to structural analysis ?

Harold TELLIER

Directeur: Prof. I. LINDEN

Mémoire présenté
en vue de l'obtention du titre de
Master 120 en ingénieur de gestion, à finalité spécialisée
en Analytics & Digital Business

ANNEE ACADEMIQUE 2020-2021

Acknowledgements

Before going into the subject of this thesis, I would like to thank several people who have been a great help in the completion of this work and my studies in general.

First of all, I would like to thank Mrs. Linden, my thesis supervisor, for her patience and flexibility. I would also like to thank her for her judicious advice and the guidance she gave me when I needed it.

I would also like to thank my family who have been a great source of motivation and support throughout my academic career and even more during the completion of this thesis. A special thank you to my brother Tom Tellier who introduced me to these studies and who helped me in the rereading of this thesis. I would also like to thank my friends who have been a great support and have given me a critical opinion on my own work.

Finally, I would like to thank the University of Namur for the quality of the teaching received, as well as its professors for their pedagogy and their availability.

Contents

1	Introduction	4
1.1	What is the structural analysis ?	4
1.2	Graph visualization	5
2	Understanding of graphs	6
2.1	Graph drawing aesthetics	6
2.1.1	Nodes heuristics	6
2.1.2	Edges heuristics	7
2.1.3	Layouts heuristics	9
2.2	Node-link layouts	10
2.2.1	Spring force-directed layout	10
2.2.2	Planar graph layout	10
2.2.3	Topological features layout	11
3	New tools to support structural analysis	12
3.1	Context	12
3.2	Methodology	13
3.2.1	Target population	13
3.2.2	Survey method	13
3.2.3	Sampling method	14
3.3	Questionnaire	15
3.3.1	Respondents characteristics and technical questions	15
3.3.2	Tool 1 : Colors	15
3.3.3	Tool 2 : Positions	16
3.3.4	Tool 3 : Shapes	16
3.4	Data preparation and analysis	17
3.5	Results analysis	18
3.5.1	Sample description	18
3.5.2	Research sub-question answers	19
3.6	Research question answers	27
4	Limitations and discussion	28
4.1	Limitations of study	28
4.2	Discussion and Future work	29
5	Conclusion	30
A	Questionnaire - Respondents' characteristics	31
B	Questionnaire - Colors	31
C	Questionnaire - Positions	33
D	Questionnaire - Shapes	35

Abstract

Today's world is a world of data, and the biggest challenge is to manage and make use of this data by structuring it in order to extract information. Therefore, information visualization, and more specifically graph visualization, is experiencing a real boom and is the subject of more and more research.

In this thesis, we are interested in structural analysis, a technique for structural and semantic analysis of content, and in the different supports that graph visualization can provide to it, particularly from the point of view of readability and interpretation. However, this last point is not really analyzed and is nevertheless essential in the context of structural analysis. This is why we decided to articulate this thesis around the following research question: "How graphical tools such as colors, shapes and positions of nodes or links can affect interpretation of graphs: support structural analysis?"

In order to answer this question, in the first part, we focused on the readability and understanding of the graphs. We have therefore tried to gather the different guidelines for creating a readable graph and explain their tradeoff in the context of graph layouts.

Then, in the second part, we discussed the question of the interpretation of graphs during a survey. We analyzed the influence of colors, shapes and positions of graphical elements on the interpretation of a graph by readers and the way they represent graphs.

1 Introduction

1.1 What is the structural analysis ?

Since the end of the 1960s and the beginning of the 1970s, many sociologists have been interested in different techniques for analyzing content representations, especially "structural analysis", in various fields.[14].

This can be considered as a semantic and structural content analysis. Indeed, in addition to analyzing the organization of the content, it is also interested in the implicit meaning given to it[14, 19].

Therefore, it has two levels of analysis. Firstly, it seeks to understand the many relationships of opposition or association within the discourse, i.e. the structure. Secondly, it seeks to understand the exact meaning of the author's discourse, what he intended to mean by his words, i.e. the meaning[14].

To achieve this, this content analysis technique is based on the postulate that the meaning of words is defined by the relations that exist between them[14], especially differences between words[19]. Thus, the meaning of a word is given as much by what it designates as by what it does not designate. Indeed, a word can have two completely different meanings depending on its opposite and therefore its context. Structural analysis is therefore based on relations called disjunctions which link a word to its opposite[19].

1.2 Graph visualization

Among the different techniques that make up information visualization, graph visualization is certainly the most researched[13]. This is due to its universality, it can be applied in many fields such as biology, chemistry, medicine, finance, software engineering and many others[5, 13, 17, 9].

This technique is mainly used to represent structured data, i.e. when the data to be represented contains internal relationships[17, 9, 13]. The purpose of graph visualization is to create a graph from a set of connected data that represents the data and their relationships in a way that facilitates understanding and perception in order to eventually extract information[5, 17].

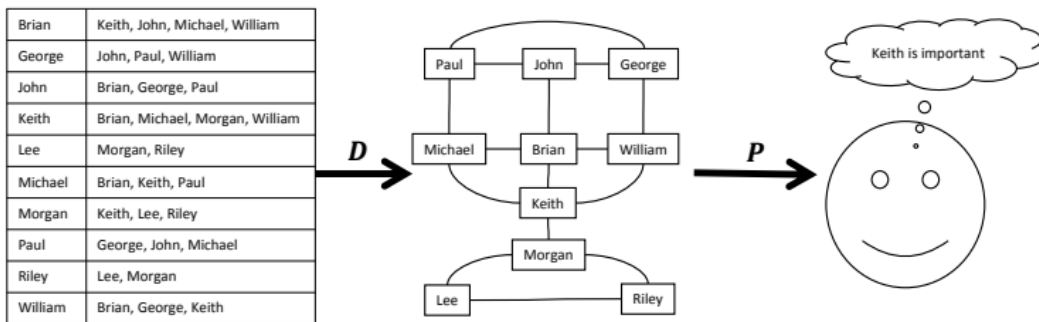


Figure 1: Graph visualization process[5]

The biggest concern in graph visualization is the readability of graphs, i.e. the ability of a graph to be correctly understood, perceived by the reader[5]. Several factors have an effect on readability, the main one being the size of the graph[17, 9]. There are criteria, called aesthetics, which help to improve the understanding of graphs.[5, 2]

This thesis is organized as follows. In section 2, we will discuss the different aesthetics that allow graphs to be more understandable and readable from the point of view of nodes, edges and overall layout as well as their practical implementation in different graph layouts. Then, section 3 will discuss the study we have carried out to understand the influence of certain graphical tools on the interpretation that the reader can make of a graph in order to find supports to structural analysis. Finally, section 4 will highlight the various limitations and obstacles encountered during this study.

2 Understanding of graphs

2.1 Graph drawing aesthetics

2.1.1 Nodes heuristics

When drawing a graph, it is customary to start by arranging the various nodes that will make up its structure. To carry out this step in the best possible way, there are several heuristics.

The first heuristic proposed for the placement of nodes is an equal distribution of vertices within the graph[17, 4]. As the name suggests, this consists of spreading the nodes equally throughout the graph space but not necessarily spacing the nodes by a uniform distance[4], i.e. according to Taylor and Rodgers with their homogeneity criterion, it must be possible to divide the graph into quadrants and each of these quadrants must contain the same number of nodes.[18] It has a real interest in the sense that it gives the graph a more regular appearance and a more attractive visual.[2]

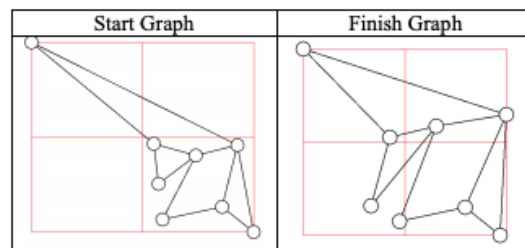


Figure 2: Taylor and Rodgers even distribution of nodes[18]

After that, there is a more semantically based node placement heuristic, the clustering of similar nodes[2]. Indeed, a study by Huang, Hong and Eades found that the reader considers that nodes close to each other belong to the same group.[10] It can therefore be very interesting to group related nodes and separate them from others in order to bring groups out[10].



Figure 3: Example of clustering nodes[22]

A third common aesthetics heuristic is the node spacing from edges which was proposed both by D. Harel and R. Davidson and consists on placing vertices at a sufficient distance from edges, i.e. "minimal distance between the node and any point on the edge"[4, 8]. This guideline is knowing the previous theory about nodes clustering. Close elements tend to be understood and interpreted by readers as similar

elements [10] and this could result in a completely misunderstanding of the graph[2].

In the same way to ensure separation between vertices is the node orthogonality[2]. The motivation behind this heuristic is to distribute nodes and (bend points) according to intersections of an imaginary Cartesian grid.[16, 2]

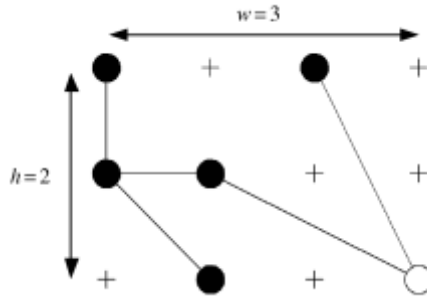


Figure 4: Node Orthogonality calculation[16]

Figure 4 gives a good example of how this aesthetic works. There are 6 nodes to distribute on a 3x2 drawing space which means that there are 12 different places, represented by "+" signs, because the number of intersections is given by $(3+1) \times (2+1) = 12$. [16]

Finally, it may seem quite obvious but non-overlapping vertices is not to be overlooked[2]. Wetherell C. and Shannon A. emphasize two points. Firstly, the labels of the nodes, which can be of very different sizes and should not overlap. Secondly, care must be taken that this overlap does not occur with labels and edges either.[21]

2.1.2 Edges heuristics

Now that the nodes are arranged on the drawing area, it is time to link them together. Once again, there are many aesthetics regarding edges, whether it is their shape, length or position.

The first guideline to be discussed is probably the one that is most widely accepted in the graph drawing literature.[2, 8, 16, 4, 18] It is the minimization of link crossings. Indeed, several studies agree that edge crossings have a real negative impact on the understanding of a graph.[18] However, this can vary depending on the angle of the crossing.[20]

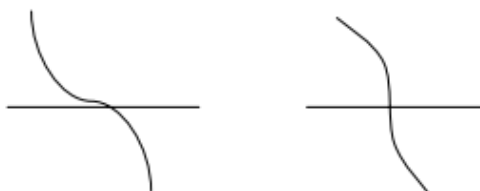


Figure 5: Difference of crossing angles[20]

On figure 5, we can see two different angles. The left graph tends to create more confusion compared to the right one because angles of its crossing are far from 90 degrees which is the less confusing.[20] Indeed, an angle that is too small or too large tends to bring the links together and make them more difficult to distinguish [18].

Another consideration is the minimization of edge bends. Bends are also a determining factor in the readability of a graph as it is not always easy to follow a link that is not straight[18] and turns around several elements, which reduces the understanding of the graph.[2, 18]

If it is not possible to avoid these bends, there is an aesthetic that advocates a certain uniformity in the bend of a link. It is therefore possible to place restrictions on the location of the bend on the edge but also on the bending of the link[2], in order to have a better uniformity and more regular graphs[18].

After that, it is important to pay attention to the length of these links. This plays an important role in the readability and comprehension of a graph. Firstly, we talk about minimizing the length of edges, which means that their size should be reduced as much as possible in order to reduce the overall space of the graph[2] but it needs to be done carefully in order not to reduce it too much, otherwise it would be too small and the aesthetic would have the exact opposite effect.[4]

Following on from this aesthetic, it is suggested to use another one: edge length uniformity[2, 18]. This allows the graph to have a more regular appearance and thus an easier visualization.[18]

Furthermore, when several edges have the same origin (start node), it is preferable to maximize the minimum angle of these edges[2, 3]. In fact, this consists in always increasing the smallest angle formed by the edges, so that they are more distinguishable[18]. If this aesthetic is perfectly respected, all angles are equal in the end[16].

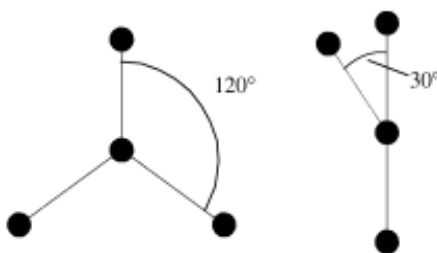


Figure 6: Maximization of minimum angle[16]

Finally, the last edge placement heuristic is the edge orthogonality maximization. It follows the same principle as for the nodes, i.e. the links are fixed to a fictitious Cartesian grid[16], which has the immediate effect of improving the understanding of the graph[15]. In addition to this, it has a positive effect on two other previous aesthetics: minimization of crossings and maximization of the smallest angle[2].

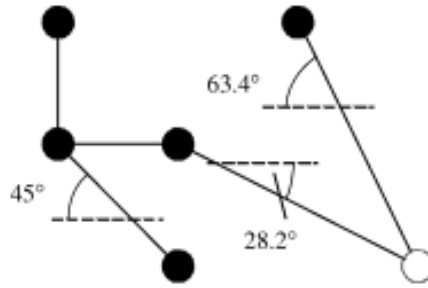


Figure 7: Edge Orthogonality[16]

2.1.3 Layouts heuristics

In addition to the guidelines about nodes and edges, there are several aesthetics about the general layout of the graph.

The first is a geometrical consideration, namely maximizing the symmetry of the graph, which has the effect of considerably increasing the readability of a graph[15]. This can be done in two ways: globally or locally.

Firstly, as its name indicates, the global symmetry aesthetic consists in arranging a graph in such a way that it is possible to divide, with an axis, this graph into two identical parts[16].

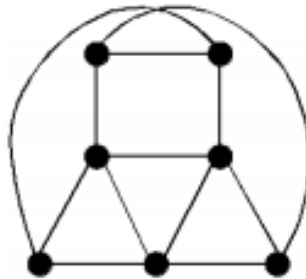


Figure 8: Global symmetry[2]

Secondly, local symmetry consists in performing an axial symmetry on a smaller part of the graph, a sub-graph[18].

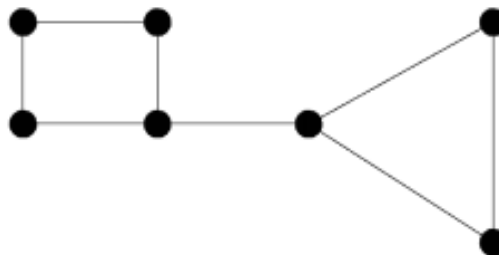


Figure 9: Local symmetry[16]

Another important factor to consider is the area occupied by the graph for which Taylor M. and Rodgers P.[18] have developed two heuristics. The first consideration is to minimize the area of the graph, while avoiding reducing the readability and comprehensibility of the graph. The second is to match the aspect ratio of the graph to that of its support. This ratio is the result of the ratio of the width and length[18]. All of this greatly increases the ease of viewing the graph.[2]

2.2 Node-link layouts

Now that we have a good overview of the different aesthetics that make graphs more readable, we will now see different types of node-links layouts algorithms that uses some of these aesthetics as a general rule for drawing understandable graphs.

2.2.1 Spring force-directed layout

Spring force-directed layout algorithms are regularly used to draw node-link diagrams.

To understand how they work, we need to use the analogy of a mechanical system[17]. In this system, nodes are replaced by steel rings and edges by springs. The latter are endowed with an attractive and a repulsive force[6]. The algorithms seek to find the arrangement of the rings that allows the energy of the system to be reduced as much as possible[12]. When the balance state is reached, the arrangement is in its final form.

These layout algorithms are very practical because it respects to aesthetics: symmetry and uniformity of the edges[12]. However, its big weakness is its instability, i.e. the same algorithm can give two different results on the same input[17].

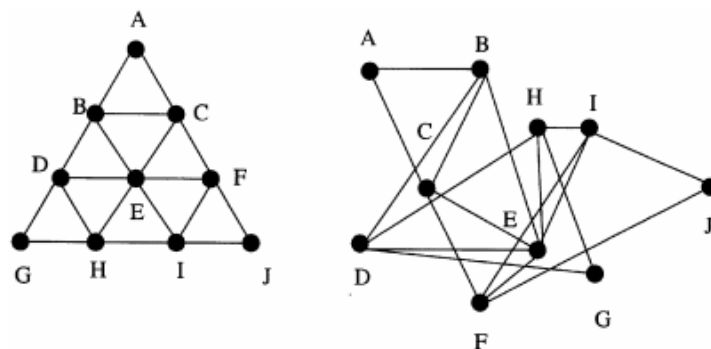


Figure 10: Left: Spring layout algorithm output graph
Left: Spring layout algorithm input graph [6]

2.2.2 Planar graph layout

Then planar graphs are probably the most common. The main consideration for this type of layout is edge crossings[17]. Indeed, planar graph algorithms create layouts

that have absolutely no edge crossing with a linear time complexity [11].

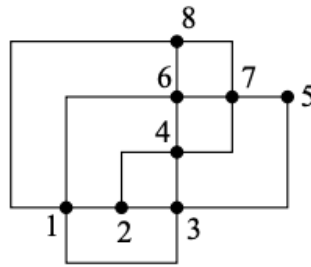


Figure 11: Example of a planar graph, more precisely an orthogonal drawing) [11]

2.2.3 Topological features layout

Finally, topological graph layout algorithms are less well known but are also a method of creating node-links diagrams.

Their process is divided into four phases. The first phase consists of recursively dividing an input graph into sub-graphs[1]. After this step, the finally found sub-graph is modified according to the feature type[17]. Then several checks and corrections are made to, firstly, eliminate crossings and, secondly, eliminate overlapping nodes[1]. We can therefore notice that two aesthetics are put forward in this type of layout.

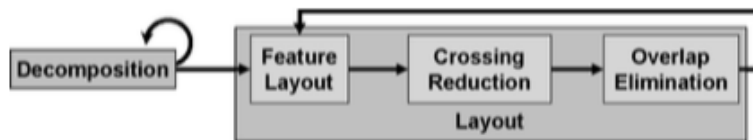


Figure 12: Topological feature-based layout algorithms process) [1]

3 New tools to support structural analysis

3.1 Context

As we saw through the exploration of the existing literature of graph visualization, there is a lack of studies about different ways to represent information towards interpretation, perception and thus understanding of the message behind the graph which is a central concern when it comes to structural analysis.

Indeed, the current literature about graph drawing set the focus on the best practices, also called aesthetics in order to have a “beautiful, good” graph in an effort to optimize its readability such as a good distribution of nodes and equal length of links but in this study, the object will be on the different graphical possibilities such as position, shape and color of nodes and links between themselves to make the information of the text behind the graph the most understandable. The study question would then be “How graphical tools such as colors, shapes and positions of nodes or links can affect interpretation of graphs to support structural analysis ?”

In order to answer this question, it will be divided into three different sub-questions:

- **How can links colors affect reader’s interpretation and understanding of nodes relation in graphs?**

In this part, the interpretation of colors will be studied. Indeed. Indeed, connected nodes can be associated or dissociated according to the color of their links. The different hypothesis tested will be further discussed and explained in the section Questionnaire : Colors

- **How can nodes positions affect readers’ interpretation and understanding of nodes relation in graphs?**

In a second part, it would be interesting to see if the position between two connected nodes can influence readers interpretation and thus if it can be understood as different types of relations. The different hypothesis tested will be further discussed and explained in the section Questionnaire : Positions

- **How can links shapes affect reader’s interpretation and understanding of nodes relation in graphs?**

Thirdly, the different forms of links could be studied. Different shape of link can be interpreted as different type of relationships in terms of intensity. The different hypothesis tested will be further discussed and explained in the section Questionnaire : Shapes

3.2 Methodology

To answer those questions and thus have a better idea on how graphs can be interpreted with these graphical tools, a survey will be conducted as quantitative study. Results can only be significant if enough respondents take part in this survey so that it is possible to draw conclusions. Once data will be collected, it will be imported into an R dataframe and statistically analyzed.[7]

3.2.1 Target population

Given that the goal of this analysis is to provide a reliable overview of the influence of these tools on the interpretation and understanding of graphs, it is important to have a very diverse population including every group or class. Indeed, the socio-cultural characteristics of the respondents will probably have a great impact on their ability to interpret. Another important criterion will be a good comprehension of the French language. As the survey aims to find support for the structural analysis of texts, it seems essential that respondents have a good comprehension of French in order to answer honestly and consciously. Moreover, this will prevent the language barrier and reduce the respondents' fear of participating in the survey[7]. It may also be useful to restrict participation to people over the age of 18 as we don't really want to draw conclusions for people under that age.

It is possible that in the context of this thesis, it is finally decided to focus on the 18-24 age group in order to avoid any representativeness bias for the other age groups. Indeed, the survey method will be more likely to reach respondents in this age group.

3.2.2 Survey method

In order to conduct this study, it was decided that the survey will be done thanks to electronic survey on Internet method[7], namely using Google Forms. For this study to be meaningful and credible, it is essential to have a large enough number of respondents to be able to infer on the complete population. Thus, this method is without doubt one of the best method as it allows to collect a decent number of answers in a relatively small time-range as it will be spread on social networks for free[7].

Although this method may lead to not very serious responses, it almost completely prevents interviewer bias by guaranteeing anonymity to the respondent and more freedom in their answers. [7]

3.2.3 Sampling method

With regard to the sampling method, given the survey method chosen, the convenience method seems the most suitable because it is free and simple. It is a non-probabilistic technique, which means that not everyone has the same probability of being chosen for the sample[7]. Indeed, as the survey is distributed on social networks, the respondents will be the people who are connected at that moment. There is therefore a risk of self-selection bias[7].

Moreover, as it is common in social network surveys, it was decided to use the snowball method as well. This method consists in the fact that respondents are given the opportunity to share the survey[7]. As a result, this survey will reach the contacts of the first respondents, then the contacts of the contacts, etc. Hence the snowball method. This technique is also free and very simple allowing to always reach more and more people.

Finally, in order to draw conclusions, it is necessary to have a sufficiently representative sample of the population. However, given the survey method chosen, it is reasonably likely that older age groups are under-represented, so the conclusions will not focus on these age groups. A lower limit on the number of respondents is set at 100 to avoid this problem as much as possible.

3.3 Questionnaire

3.3.1 Respondents characteristics and technical questions

In the first part, respondents will be asked to answer questions mainly about their characteristics (See Appendix A). This step is essential because it will allow the responses to be analyzed from different angles and will make it possible to draw conclusions according to the different types of respondents. In addition, it will ensure that the respondents correspond well to the target population, as they will be asked if they have a sufficient knowledge of the French language at the beginning of the questionnaire, for the reasons explained in the previous subsections. Responses that do not meet this condition will not be analyzed and removed before analysis. Several personal characteristics will therefore be requested, such as age, sex, level of education and type of profession. This information will, of course, be kept and declared confidential to not repel respondents.

Then the specific questions at the heart of this survey will be asked. In order to find out more about the different possible interpretations of the selected graphic tools, respondents will be asked several types of questions. The first type is to see how the respondents graphically represent a text composed of several relations of opposition, similarity, etc. To do this, series of response graphs have been created beforehand, some highlighting certain assumptions, others none. Secondly, the respondents will be asked to perform the opposite operation, i.e. they will have to choose a text which, according to them, best corresponds to the graph from a list also established beforehand. These two questions will allow for an honest and "naive" opinion. Finally, respondents will be asked to choose the type of relationship represented by a graphical relation without any context.

These will be divided into three parts according to the theme of the questions (Animals, People or Smartphones). The only purpose of this separation is to confuse the respondents and thus limit as much as possible the feeling of desired answers.

3.3.2 Tool 1 : Colors

In this part of the questionnaire, we get to the heart of the matter. It will allow us to answer or not the question "How can links colors affect reader's interpretation and understanding of nodes relation in graphs? To do this, several assumptions will be investigated:

- H1 : The green color reflects a membership/similarity relationship between two concepts.
- H2 : The color red reflects an opposition/difference relationship between two concepts.

3.3.3 Tool 2 : Positions

The second set of questions will focus on "How can nodes positions affect readers' interpretation and understanding of nodes relation in graphs?". It aims to understand if any difference of nodes position can be interpreted in different ways such as these hypotheses.

- H3 : A difference in horizontal position reflects a equivalence/similarity between two connected nodes.
- H4 : A difference in vertical position reflects a superiority/inferiority relationship between two connected nodes.

3.3.4 Tool 3 : Shapes

Finally, this part will try to help address the last sub-question "How can links shapes affect reader's interpretation and understanding of nodes relation in graphs?" in order to see if nuances in relation can be supported by different shapes of links.

- H6 : A discontinuous link reflects a low intensity relationship.
- H7 : A continuous link reflects a relationship of medium or high intensity relationship.

3.4 Data preparation and analysis

First, as the survey program used does not offer the functionality required to analyze the data, the data will be exported as an excel file and imported into a program that uses the R programming language and allows for more complex analyzes (R Studio). Thanks to this, the data can be stored in a data structure, called dataframe, and analyzed with the different packages offered by R such as ggplot, dplyr, etc.

Then, each question will be analyzed individually in relation to the different characteristics of the respondents and their frequency of appearance in the sample to investigate the differences in perception and interpretation according to these parameters.

After that, it will be interesting to analyze the possible presence of dependencies between the characteristics and response variables. To do this, several bivariate analysis will be performed, such as Chi-squared and Fisher.

Finally, all those analyses will allow to verify or refute, as best as possible, different assumptions made earlier. It will probably be necessary to nuance the answers according to the type of questions asked. Indeed, it is possible that text-graph and graph-text questions will provide completely different results. It might also be interesting to nuance these with the questions specific to the hypotheses themselves.

3.5 Results analysis

3.5.1 Sample description

Over a period of thirteen days, from 27 July to 9 August, we collected responses and opinions from French-speaking people about the different possible interpretations of graphs with various graphical variants. The number of respondents to this survey was 156, including two who were removed because they did not meet the characteristics of the target population (good French language knowledge). To obtain a varied sample in terms of demographics, the survey was shared on social networks so that all groups of characteristics could be sufficiently represented, but it turned out that not all groups were sufficiently represented and that the sample is not sufficiently representative to perform analysis on all age groups in the population, as announced in the definition of the target population.

Indeed, when we look at the proportions of respondents by age, we can notice that the three oldest age groups are clearly under-represented with a total of 12.8% for those aged 35 to over 55. The largest group is therefore the 18-24 year old people with 62.2%. This is largely due to the chosen survey method, i.e. the online survey on social networks. As a result, there are also considerable differences in shares of profession. More young people means more students, and by the same rationale, fewer older people means fewer pensioners. Thus, 42.9% are students, 31.4% are employees, 9.6% are self-employed, 8.3% are doing a liberal profession. The last three categories account for 7.6%.

However, to avoid representativeness bias as much as possible, the following analysis will only be carried out on the 18-24 age group, which we will refer to as respondents or 18-24 years olds from now on.

So within this sub-sample, there is a surprisingly even gender distribution, with 52.6% men and 47.4% women. This is fairly representative of the population.

Then, the sub-sample consists of 65.3% students, 22.1% employees, 6.3% self-employed, 3.1% liberal profession practitioners, 1.1% unemployed, 2.1% laborer's which is more representative of the population of this age group.

Finally, the distribution of education level offers a great disparity as there are simply no respondents whose maximum level of education is primary education, 12.8% secondary education and 87.2% higher education. This can be explained by the fact that the survey was widely shared via a Facebook group of students from the Namur.

3.5.2 Research sub-question answers

Question 1 : How can links colors affect reader's interpretation and understanding of nodes relation in graphs?

In this subsection, we will analyze the results of the survey, focusing on those related to the influence of colors on the interpretation of 18-24 year olds. Once this is done, we will try to draw conclusions from the different results and statistically analyze them in order to answer the first research sub-question.

Before starting these analyses, let's briefly review the different hypotheses we wanted to test related to colors:

- H1 : The green color reflects a membership/similarity relationship between two concepts.
- H2 : The color red reflects an opposition/difference relationship between two concepts.

First, we will analyze the impact of colors on the representational skills of 18-24 year olds, i.e. how they graphically represent relations of association and opposition. The question 1 (See Appendix B.1) was to select a representation for a text including previously mentioned relations. As we can see on the graph below, 48.4% of respondents chose the response "H" which was the graph respecting both assumptions H1 and H2. The response with the second highest frequency was response "G" (30.5%) which represented only hypothesis H2, followed by response "D" showing only hypothesis H1 with 11.6%.

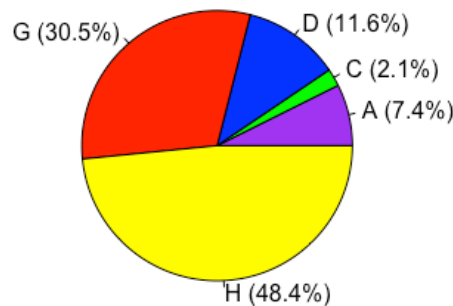


Figure 13: Representations of association and opposition relationships

After that, potential dependencies between the respondents' personal characteristics and their interpretations will be investigated. To check this, two statistical tests will be carried out, Chi-squared and Fisher (with a p-value of 0.05, thus with a reliability level of 95%). Only the gender and profession variables will be tested, as the education variable is not sufficiently representative.

At first sight, the dependency tests are not conclusive. Indeed, the p-values obtained with these tests are lower than 0.05. Hence, it can be seen that there is no significant dependency between the representations chosen by the respondents and

their personal characteristics, either for gender or for profession.

Next, we will look at the respondents' ability to interpret colors, i.e. how they interpret colored links into text form. Question 2 (See Appendix B.2) asked respondents to choose the text that best fits the graph. The results of this question are even more obvious and equivocal, as shown in the figure below. There is an overwhelming majority of 87.4% for answer B, which was the only answer that interpreted all green links as similarities and all red links as differences.

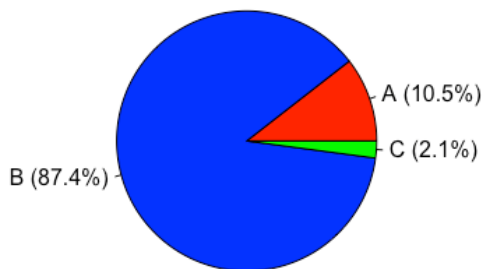


Figure 14: Interpretations of colored links

As with the representation question, tests were carried out to learn more about the influence of personal characteristics and differences in interpretation. The finding remains the same as for the previous question, no significant dependency between the answers given to the question and the respondent's profession or gender can be established.

Finally, we will analyze the results of the last question (See Appendix B.3) which aimed to find out more about the context-free interpretation of the different links in our hypotheses as well as to verify the results of the previous questions.

The results of this question show that, without any particular context, almost 80% of the selected panel believe that a green link corresponds to a similarity relationship and just under 90% interpret a red link as the opposition of two nodes. Moreover, these results are in line with those collected in the previous questions.

	Similarity/Association	Difference/Opposition	Neither of these	Total
Green link	79% (75)	1% (1)	20% (19)	100% (95)
Red link	2.1% (2)	87.4% (83)	10.5% (10)	100% (95)
Black link	26.3% (25)	19% (18)	54.7% (52)	100% (95)

Figure 15: Representations of association and opposition relationships

Then, we will look for possible dependencies between these answers and the respondents' gender/occupation, by performing Chi-squared and Fisher tests.

Results show that the respondent's profession has an influence on his or her interpretation of the red links. Indeed, the dependency test reveals that there is a significant dependency (p-value test < 0.05) between the responses and the profession variable.

In summary, several conclusions can be drawn from these questions:

- People from 18 to 24 year old graphically represent associations with green links and oppositions with red links
- People from 18 to 24 year old interpret green links as associations and red links as oppositions
- The profession of an 18-24 year old has an impact on how he or she interprets red links

Question 2 : How can nodes positions affect reader’s interpretation and understanding of nodes relation in graphs?

In this second subsection, we will analyze the results of the survey and tackle the influence of positions on the interpretation of 18-24 year olds. Once this is done, we will try to draw conclusions from the different results and statistically analyze them to answer the second research sub-question.

First of all, we are going to make a quick reminder about assumptions investigated in the following part:

- H3 : A difference in horizontal position reflects a equivalence/similarity between two connected nodes.
- H4 : A difference in vertical position reflects a superiority/inferiority relationship between two connected nodes.

Now, we will analyze the influence of positions on the the way of representing superiority/inferiority and equivalence in graphs of 18-24 year olds, i.e. how they graphically represent nodes that are superior or inferior, and equivalent to each other. The first question (See Appendix C.1) consists of choosing a representation that best matches a text including previously mentioned relations. The results show that two answers stand out. On the one hand, answer B with 44.2%, respects the two hypotheses defined previously. On the other hand, answer F with 28.4%, respects the first hypothesis but represents superiority/inferiority by a slanting link.

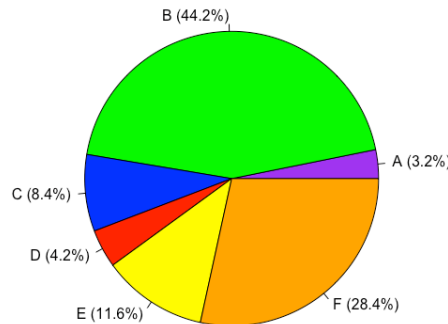


Figure 16: Representations of superiority/inferiority and equivalence relationships

To find out whether the choice of representations is affected by the personal characteristics of the selected panel, we will carry out Chi-Squared and Fisher dependency tests with a 95% level of reliability, as in the case of the tests on colors.

These tests show that it is not possible to establish that there is a significant dependency between the answers given to this question and the respondents’ personal variables, i.e. gender and occupation, because the p-values resulting from the bivariate analysis are lower than 0.05.

Next, we are interested in the interpretation of these differences in node positions within a graph. The second question (See Appendix C.2) is the reverse of the previous one. The respondent must now match a text-interpretation to a given

graph. Results show on the graph below that 54.7% of respondents answered A which means that they interpreted vertical links as superiority/inferiority relationships and horizontal links as equivalence relationships.

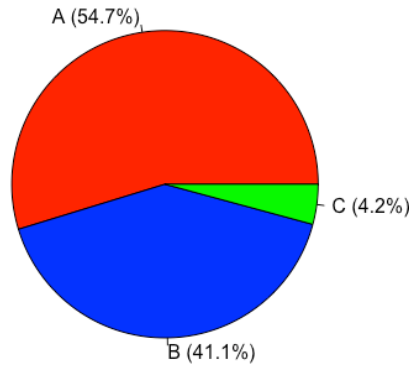


Figure 17: Interpretations of links varying in position

Finally, the last question (See Appendix C.3) will allow us to test the differences in the context-free interpretation of the relationships defined in our hypotheses. We can also compare the results with the previous questions to confirm our findings.

We can see that, in the 18-24 age group, below 90% of respondents believe that a purely vertical position difference between two connected nodes represents a superiority/inferiority relationship, and that almost 75% of respondents believe that a purely horizontal position difference between two connected nodes represents an equivalency/similarity. These results are in line with our previous findings.

	Inferiority/Superiority	Equivalence	Neither of these	Total
Vertically interconnected nodes	88.4% (84)	0% (0)	11.6% (11)	100% (95)
Horizontally interconnected nodes	0% (0)	74.7% (71)	25.3% (24)	100% (95)

Figure 18: Frequencies for context-free interpretation of links varying in position

Then, we will seek for potential dependencies between these answers and the respondents' gender/profession, by testing it through Chi-squared and Fisher analysis.

The p-values obtained thanks to these tests allow us to establish that there is a significant dependency between the occupation and the context-free interpretation of positions. Indeed, it seems that in general students are more likely to consider links between vertically aligned nodes as superiority/inferiority relations and links between horizontally aligned nodes as equivalence relations.

In summary, several conclusions can be drawn from these questions:

- People from 18 to 24 year old graphically represent superiority/inferiority with

vertically aligned connected nodes and equivalence/similitude with horizontally aligned connected nodes.

- People from 18 to 24 year old interpret horizontally aligned connected nodes as equivalence/similitude and vertically aligned connected nodes as superiority/inferiority.
- The profession of an 18-24 year old has an impact on how he or she interprets vertically and horizontally connected nodes positions .

Question 3 : How can links shapes affect reader's interpretation and understanding of nodes relation in graphs?

In this last subsection, we will examine the results of the survey and approach the influence of shapes on the interpretation of 18-24 year olds. Once this is done, we will try to draw conclusions from the different results and statistically analyze them to answer the second research sub-question.

Let's start with a reminder of the assumptions that will be investigated in this section:

- H6 : A discontinuous link reflects a low intensity relationship.
- H7 : A continuous link reflects a relationship of medium or high intensity relationship.

We are now going to analyze the impact of different shapes on the way 18-24 year olds represent intensity, i.e. how they graphically represent different intensities in relationships. The question 1 (See Appendix D.1) asked to select a representation for a text including previously mentioned relations. The next figure shows that a huge majority of 85.3% of 18-24 year olds chose graph A that represents weak relationships using discontinuous links and strong relationships using continuous links, i.e. both assumptions. The other 14.7% seem to make no difference between weak and medium/strong relationships.

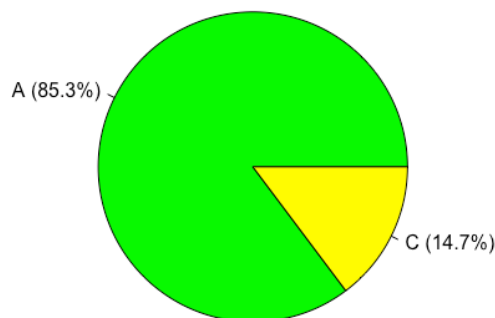


Figure 19: Representations of weak and medium/strong relationships

As usual, we will now carry out different dependency tests (Chi-Squared and Fisher) with a reliability threshold of 0.05 and thus a reliability level of 95%, to find out whether gender or occupation has any influence on the way in which differences in intensity in the links are represented.

None of the tests for any of the variables show a result below this threshold. It can therefore not be concluded that these characteristics have an influence on the chosen representation.

After that, we will analyze how 18-24 years olds interpret graphs containing links of different shapes, continuous and discontinuous with question 3 (See Appendix D.2). It can be noted that proposition D, which is the one that interprets all the links in concordance with our hypotheses, is the most selected answer with

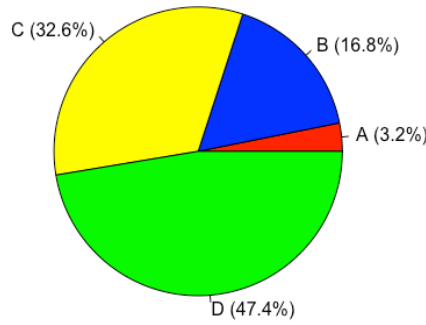


Figure 20: Interpretations of links shapes

47.4%, a little less than 1 in 2 people.

After performing Chi-Squared and Fisher dependency tests with personal characteristics, it appears that the gender of the respondent would influence the interpretation of a graph consisting of links of different intensity. In general, women are more likely to choose the interpretation that matches discontinuous links to weak relationships and continuous links to medium and strong relationships.

At last, we will analyze the results of the very last question (See Appendix D.3) which aimed to find out more about the context-free interpretation of the different shapes of links and maybe allow us to confirm previous results. Frequencies give clear results, 3 out of 4 people consider a discontinuous link to be a low intensity relationship between two nodes and a continuous link is considered a high intensity relationship by 72.6% and a medium intensity relationship by 26.3% which confirms our results and assumptions.

	Null	Weak	Medium	Strong	Total
Discontinuous link	3.2% (3)	75.8% (72)	18.9% (18)	2.1% (2)	100% (95)
Continuous link	0% (0)	1.1% (1)	26.3% (25)	72.6% (69)	100% (95)

Figure 21: Frequencies for context-free interpretation of links shapes

We will now perform the last test of dependency between context-free interpretations and personal characteristics. Results are not significantly conclusive, no dependency can be concluded.

In summary, several conclusions can be drawn from these questions:

- People from 18 to 24 year old graphically represent weak relationships with discontinuous links and medium/strong relationships with continuous links
- People from 18 to 24 year old interpret weak relationships as weak relationships and continuous links as medium/strong relationships.
- The respondent's gender has an influence on his or her interpretation of the shapes of links in a graph.

3.6 Research question answers

Thanks to findings extracted from analyses of the survey data and the answers to the research sub-questions, we can try to answer the initial research question "How graphical tools such as colors, shapes and positions of nodes and links can affect interpretation of graphs: support structural analysis?" With this question, we wanted to know more about the different interpretations that can be made when a link is of a certain color or shape, etc to find out more about its applicability to a structural analysis context. It seems that the three tools have a significant impact on people in the 18-24 age group.

Indeed, young people are very receptive to colors. Whether in the sense of graphical representation of a text or interpretation of a graph in the form of text, they very clearly consider green links as relations of association/similarity and red links as relations of opposition/difference. This tool is of real interest for structural analysis because this technique bases its analysis on the differences and similarities between words within a discourse. This would allow a quick and intuitive visualization for disjunctions.

Then, in line with the color tool, young people naturally interpret discontinuous links as being any relationship but of low intensity, while continuous links are interpreted as relationships of medium or high intensity. Structural analysis could benefit from this tool as it allows for the representation of a nuance in a link and thus for an even more precise grasp of the meaning of the discourse.

As for the influence of positions, the two hypotheses defined at the beginning of the study were also validated for 18-24 year olds. Indeed, the difference in vertical position between two nodes is in most cases assimilated to a complementary superiority/inferiority relationship. Conversely, a difference in horizontal position between two nodes is assimilated to a relationship of equivalence or similarity. This is probably the one of the three that seems least relevant to structural analysis. Indeed, as it stands, structural analysis is only concerned with associations and oppositions between words but not really with the sense in which they differ (one word superior to another)

It is obvious that these tools can be combined and thus used at the same time, a red discontinuous link would then be interpreted as a weak opposition relationship.

4 Limitations and discussion

4.1 Limitations of study

Through this quantitative study, we were able to learn a lot about the interpretation and perception of several graphic tools and this allowed us to draw conclusions about the 18-24 age group. However, it is essential to recognize the various limitations that affect the validity of this study from both an internal and an external perspective.

From an external point of view, the most important limitation of this study is undoubtedly the sampling procedure. Firstly, the chosen sampling method, i.e. the convenience sample, is a real obstacle to external validity as it is a non-probabilistic method. Secondly, the survey method chosen, i.e. the online survey on social networks, is generally very useful because it allows a decent number of respondents to be collected in a short space of time and at zero cost, but it often induces a self-selection bias. Indeed, this method reaches more certain age groups, mainly young people, and strongly neglects other groups, such as the elderly, who are obviously less present. Moreover, a survey on social networks often results in obtaining respondents similar to the interviewer, i.e. people of the same age or social group. This is known as over-representation and reduces external validity.

With regard to internal validity, although we have tried to limit this risk as much as possible by trying not to create a questionnaire that takes too long to answer and by guaranteeing anonymity, it is possible that with the survey method, the respondent is not necessarily honest in his or her answers, which would make the results erroneous. In addition to this, one should be wary of a potential halo effect. To mitigate this as much as possible, we 'mixed' the questions by context to avoid leading questions and response contagion.

4.2 Discussion and Future work

In the previous section we have analyzed the results of our survey and tried to establish various conclusions that are in the framework of a thesis quite relevant and representative of the 18-24 year old population. The reader now has all the cards in hand to create a graph that is understandable thanks to the aesthetics described in section 2 and meaningful thanks to the conducted study.

First of all, with regard to the study itself, it would be very interesting to dig deeper into the research. On the one hand, by replicating this study on a more representative sample of the population, thus much more diversified in terms of personal characteristics, so that all types of occupation and level of education are sufficiently represented. Secondly, by replicating the study on a larger sample to increase the external validity of the results. Moreover, it might also be worth investigating other personal factors that might influence the perception and interpretation of these graphical tools, such as ethnicity, religious beliefs, type of studies done, to find out differences between individuals.

Furthermore, in the context of this thesis related to structural analysis, we have limited ourselves to association and opposition relationships as these are the only relationships used in the context of this content analysis technique. However, this study could open the door to much more research into the influence of these tools on the interpretation of graphs by testing assumptions on other types of links applicable in other contexts. For example, it could be tested if in a directed graph, a green link is interpreted as a positive causality.

5 Conclusion

The purpose of this thesis was to understand how several graphical tools could be interpreted to make a graph that is both understandable and meaningful so that it can be used to support structural analysis. First, the main themes of this work, structural analysis and graph visualization, were briefly introduced.

Then, we focused on the understanding of graphs. We have thus addressed the subject of aesthetics of graph drawing which are guidelines concerning the placement of nodes, edges and the overall display of these elements together with the objective to obtain a readable and therefore understandable graph.

Then, for the question of meaningfulness, we asked ourselves and conducted an investigation to answer "How graphical tools such as colors, shapes and positions of nodes or links can affect interpretation of graphs to support structural analysis?"

It was found that the three analyzed tools do influence the graphical interpretations of 18-24 year olds such that for colors, green and red links are interpreted as associations and oppositions respectively. For positions, we can conclude that two vertically aligned nodes vertically induce a superiority/inferiority relationship between them, whereas two horizontally aligned nodes induce an equivalence relationship between them. Finally, for shapes, this survey showed that a discontinuous link is interpreted as a weak relationship and a continuous link as a medium/strong relationship.

Finally we tried to answer our research question by briefly explaining how the analyzed tools could support the structural analysis. It emerged that the colours and shapes as analyzed were very good tools to represent the different disjunctions that link a word to its opposite and their relations. With regard to positions, the different hypotheses defined do not seem sufficiently relevant in the context of structural analysis.

A Questionnaire - Respondents' characteristics


Formulation of the question	Possible answers (only single answer)
Do you have a good knowledge of French language?	1) Yes 2) No
In which age group do you fit?	1) 18-24 2) 25-34 3) 35-44 4) 45-54 5) 55 or more
Quel est votre sexe ?	1) Male 2) Female
What is your level of education?	1) Primary school 2) Secondary school 3) Higher school
What is your profession?	1) Student 2) Employee 3) Self-employed 4) Liberal profession 5) Unemployed 6) Laborer 7) Retired

B Questionnaire - Colors



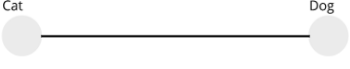
Question B.1

Formulation of the question	Possible answers
<p>Choose the graph that you think best fits the following text:</p> <p>"Dogs and cats are very common pets. They behave very differently from the mouse."</p>	

Question B.2

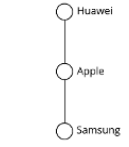
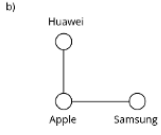
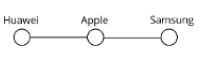
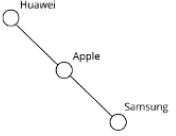
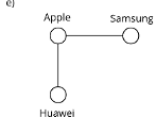
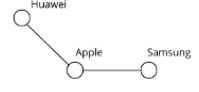
Formulation of the question	Possible answers
<p>Choose the proposition that you think best fits the following graph.</p> 	<p>One single answer</p> <ol style="list-style-type: none"> 1. The cat and the dog are very common pets, but the elephant and the lion are not. 2. Unlike the dog, the lion's behaviour is similar to that of the cat. Elephants and lions live in the savannah. 3. Elephants and lions have different diets. The cat, like the dog, is a domestic animal, unlike the lion

Question B.3

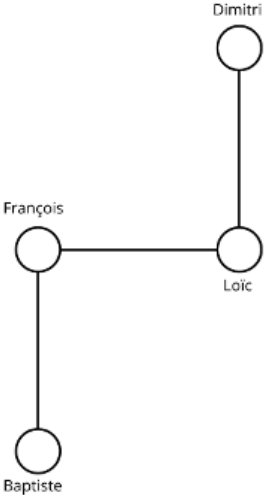
Formulation of the question	Possible answers
<p>How would you describe the relationship between "Cat" and "Dog" as represented on the graph?</p> <p>A</p>  <p>B</p>  <p>C</p> 	<p>One single answer</p> <ol style="list-style-type: none"> 1. Similarity/Association 2. Difference/Opposition 3. Neither of these

C Questionnaire - Positions



Question C.1

Formulation of the question	Possible answers
<p>Choose the graph that you think best fits the following text:</p> <p>"Huawei mobile phones are cheaper than Apple phones. They have a battery comparable to that of Samsung".</p>	<p>One single answer</p> <div style="display: flex; flex-wrap: wrap;"> <div style="width: 33%; text-align: center;"> <p>a)</p>  </div> <div style="width: 33%; text-align: center;"> <p>b)</p>  </div> <div style="width: 33%; text-align: center;"> <p>c)</p>  </div> <div style="width: 33%; text-align: center;"> <p>d)</p>  </div> <div style="width: 33%; text-align: center;"> <p>e)</p>  </div> <div style="width: 33%; text-align: center;"> <p>f)</p>  </div> </div>

Question C.2

Formulation of the question	Possible answers
<p>Choose the proposition that you think best fits the following graph.</p> <div style="text-align: center;">  </div>	<p>One single answer</p> <ol style="list-style-type: none"> 1. The best student in the class is Dimitri. On the other hand, Baptiste is the worst. As for François and Loïc, they are average students. 2. Dimitri is a good student, he is better in maths than Loïc. Loïc is also strong in chemistry, while François has more facility than Baptiste in physics. 3. The ranking of the students according to their average is as follows: Dimitri first, Loïc second, François third and Baptiste last

Question C.3

Formulation of the question	Possible answers
<p data-bbox="264 331 909 392">How would you describe the relationship between "Huawei" and "Apple" as represented on the graph?</p> <p data-bbox="427 622 446 649">A</p>  <p data-bbox="427 846 446 873">B</p> 	<p data-bbox="1010 338 1173 365">One single answer</p> <ol data-bbox="1010 394 1220 533" style="list-style-type: none"><li data-bbox="1010 394 1220 421">1. Superiority/Inferiority<li data-bbox="1010 450 1220 477">2. Equivalence<li data-bbox="1010 506 1220 533">3. Neither of these

D Questionnaire - Shapes



Question D.1

Formulation of the question	Possible answers
<p>Choose the graph that you think best fits the following text:</p> <p>"Dogs and cats are very common pets. They have slightly different behaviours compared to the mice."</p>	<p>One single answer</p> <div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p>a)</p> </div> <div style="text-align: center;"> <p>b)</p> </div> </div> <div style="display: flex; justify-content: space-around; margin-top: 20px;"> <div style="text-align: center;"> <p>c)</p> </div> <div style="text-align: center;"> <p>d)</p> </div> </div>

Question D.2

Formulation of the question	Possible answers
<div style="text-align: center; margin-top: 50px;"> </div>	<p>One single answer</p> <ol style="list-style-type: none"> 1. Nicolas is much stronger than François at chess, who himself is much better than Jean and Ludovic at football. 2. Ludovic is a very poor student compared to François, who has slightly better results than Jean and Nicolas. 3. Jean is slightly better than François at football. Nicolas is slightly less intelligent than him but Ludovic is almost as good as François. 4. François is a good student, he has much better results than Nicolas and Jean but slightly worse marks than Ludovic.

Question D.3

Formulation of the question	Possible answers
<p data-bbox="264 331 1008 389">How would you describe the relationship between "Dimitri" and "Jean" as represented on the graph?</p> <p data-bbox="336 622 355 645">A</p>  <p data-bbox="336 786 355 808">B</p> 	<p data-bbox="1008 331 1327 360">One single answer</p> <ol data-bbox="1008 389 1327 584" style="list-style-type: none"><li data-bbox="1008 389 1327 418">1. Null<li data-bbox="1008 445 1327 474">2. Weak<li data-bbox="1008 501 1327 530">3. Medium<li data-bbox="1008 557 1327 586">4. Strong

References

- [1] Daniel Archambault, Tamara Munzner, David Auber, et al. Topolayout: Graph layout by topological features. *IEEE Information Visualization Posters Compendium (InfoVis' 05)*, pages 3–4, 2005.
- [2] Chris Bennett, Jody Ryall, Leo Spalteholz, and Amy Gooch. The aesthetics of graph visualization. In *CAe*, pages 57–64, 2007.
- [3] Isabel F Cruz and Roberto Tamassia. Graph drawing tutorial. URL: *www.cs.brown.edu/rt/papers/gd-tutorial/gd-constraints.pdf*, 1998.
- [4] Ron Davidson and David Harel. Drawing graphs nicely using simulated annealing. *ACM Transactions on Graphics (TOG)*, 15(4):301–331, 1996.
- [5] Peter Eades and Karsten Klein. Graph visualization. In *Graph Data Management*, pages 33–70. Springer, 2018.
- [6] Peter Eades and Xuemin Lin. Spring algorithms and symmetry. *Theoretical Computer Science*, 240(2):379–405, 2000.
- [7] Wafa Hammedi. Course : Etude de marché. 2017.
- [8] David Harel. On the aesthetics of diagrams. In *International Conference on Mathematics of Program Construction*, pages 1–5. Springer, 1998.
- [9] Ivan Herman, Guy Melançon, and M Scott Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on visualization and computer graphics*, 6(1):24–43, 2000.
- [10] Weidong Huang, Peter Eades, and Seok-Hee Hong. *Layout effects: Comparison of sociogram drawing conventions*. Citeseer, 2005.
- [11] Goossen Kant. *Algorithms for drawing planar graphs*. PhD thesis, 1993.
- [12] Wanchun Li, Peter Eades, and Nikola Nikolov. Using spring algorithms to remove node overlapping. *APVis*, 45:131–140, 2005.
- [13] M Scott Marshall, Ivan Herman, and Guy Melançon. An object-oriented design for graph visualization. *Software: Practice and Experience*, 31(8):739–756, 2001.
- [14] Anne Piret, Jean Nizet, and Etienne Bourgeois. L’analyse structurale. une méthode d’analyse de contenu pour les sciences humaines, bruxelles, de boeck, 173p. 1996.
- [15] Helen Purchase. Which aesthetic has the greatest effect on human understanding? In *International Symposium on Graph Drawing*, pages 248–261. Springer, 1997.
- [16] Helen C Purchase. Metrics for graph drawing aesthetics. *Journal of Visual Languages & Computing*, 13(5):501–516, 2002.
- [17] Raga’ad M Tarawaneh, Patric Keller, and Achim Ebert. A general introduction to graph visualization techniques. In *Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering-Proceedings of IRTG 1131 Workshop 2011*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.

- [18] Martyn Taylor and Peter Rodgers. Applying graphical design techniques to graph visualisation. In *Ninth International Conference on Information Visualisation (IV'05)*, pages 651–656. IEEE, 2005.
- [19] Anne Wallemacq, Jean-Marie Jacques, and Vincent Bruyninckx. *Dans le sillage des mots...: EVOQ. Logiciel de cartographie cognitive*. Presses universitaires de Namur, 2004.
- [20] Colin Ware, Helen Purchase, Linda Colpoys, and Matthew McGill. Cognitive measurements of graph aesthetics. *Information visualization*, 1(2):103–110, 2002.
- [21] Charles Wetherell and Alfred Shannon. Tidy drawings of trees. *IEEE Transactions on software Engineering*, (5):514–520, 1979.
- [22] Scott White and Padhraic Smyth. A spectral clustering approach to finding communities in graphs. In *Proceedings of the 2005 SIAM international conference on data mining*, pages 274–285. SIAM, 2005.