

**School of Information Technologies**

Faculty of Engineering & IT

**ASSIGNMENT/PROJECT COVERSHEET - INDIVIDUAL ASSESSMENT**

**Unit of Study: COMP5048**

**Assignment name: Visual Analytics (Assignment 1)**

**Tutorial time: Thursday 8-9pm Tutor name: Amyra Meidiana**

**DECLARATION**

We the undersigned declare that we have read and understood the *University of Sydney Academic Dishonesty and Plagiarism in Coursework Policy*, an, and except where specifically acknowledged, the work contained in this assignment/project is our own work and has not been copied from other sources or been previously submitted for award or assessment.

We understand that failure to comply with the Academic Dishonesty and Plagiarism in Coursework Policy can lead to severe penalties as outlined under Chapter 8 of the University of Sydney By-Law 1999 (as amended). These penalties may be imposed in cases where any significant portion of my submitted work has been copied without proper acknowledgement from other sources, including published works, the internet, existing programs, the work of other students, or work previously submitted for other awards or assessments.

We realise that we may be asked to identify those portions of the work contributed by each of us and required to demonstrate our individual knowledge of the relevant material by answering oral questions or by undertaking supplementary work, either written or in the laboratory, in order to arrive at the final assessment mark.

**Student ID: 490278701**

**Student Name: Mohammed Saif**

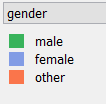
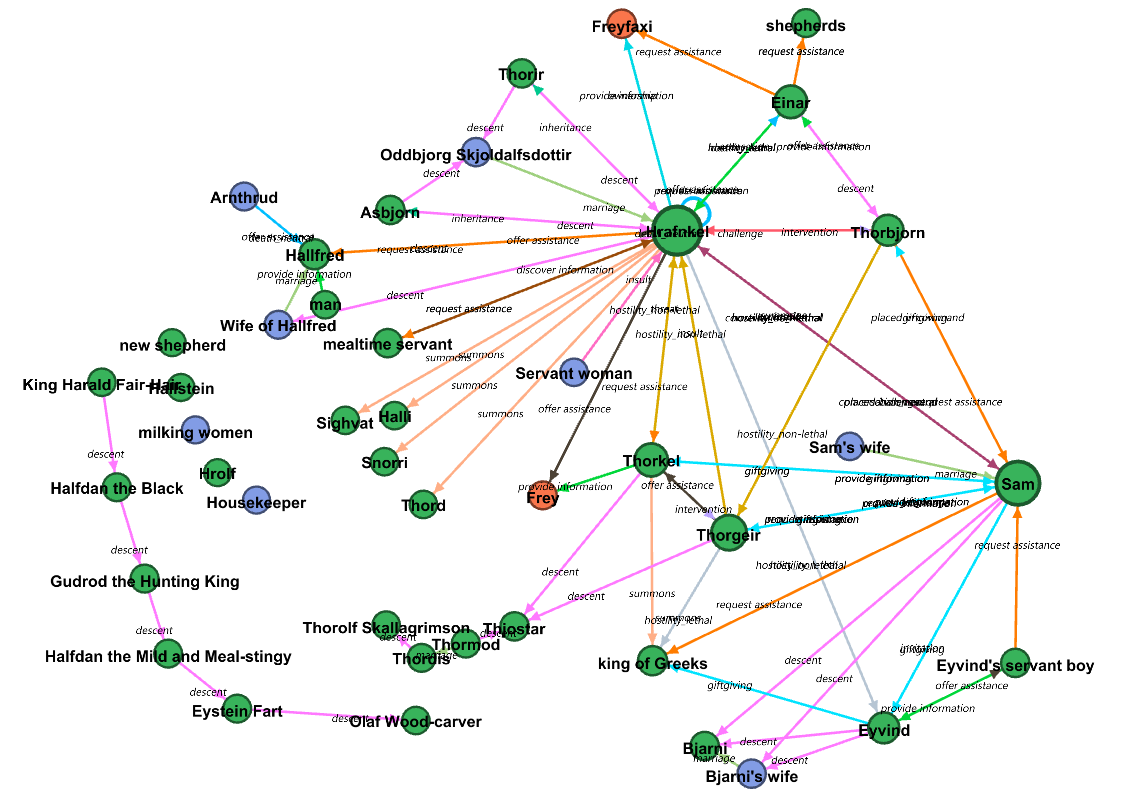
**Signed: A picture containing drawing

Description automatically generated**

**Date: 12/10/2020**

|  |  |  |
| --- | --- | --- |
| SIT Building, J12  The University of Sydney  NSW 2006 Australia | **T** +61 2 9351 3423  **F** +61 2 9351 3838  **E** [sit.info@sydney.edu.au](mailto:sit.info@sydney.edu.au)  **sydney.edu.au/it** | ABN 15 211513464 CRICOS 00026A |

**Hrafnkel Dataset**



**Node colours**

**Edge colours**

**Tools and Layouts:**

Pre-processing tool – Microsoft Excel.

The dataset had values for action (represented by the relationship between nodes) and gender in Integer. The integers were replaced by their labels in order to improve readability.

Visualization tool - Gephi.

**Implementation**

Graph algorithms used - Force Atlas + Expansion + Label adjust + NoOverlap + Manual Adjustment.

This dataset size is relatively small, therefore it could easily be visualized by Force Atlas. The computation was quick and efficient. Expansion was added to increase distance between nodes. Label adjust and NoOverlap stages were run to reduce overlapping. The nodes were also manually adjusted to fit in the same frame.

Node sizes are determined by the degree of edges (larger nodes indicate larger number of relationships - more prominent characters) and edge colours depict the relationships between the actants of the saga.

**Visual Analysis**

Hrafnkels saga is an Icelandic saga from the 10th century and depicts his relationships with different chieftains and farmers in east of Iceland. First glance indicates that Hrafnkel and Sam are the most prominent actants in the saga and also share a relationship. There are sub-components that indicate they did not interact with any actants in the saga. The most notable cluster is an ancestor-descendant chain with 6 green nodes.

Multiple labels indicate the actants had multiple types of interactions between them. For example the relationship between Hrafnkel and Sam has 3 action codes in the dataset – “place in command”, “challenge” and “conversation\_neutral”.

The node colours depend upon the gender of the actants. Other genders here can either mean the node represents an inanimate object (Freyfaxi) or a God of the saga (Freyr).

**Self-Analysis**

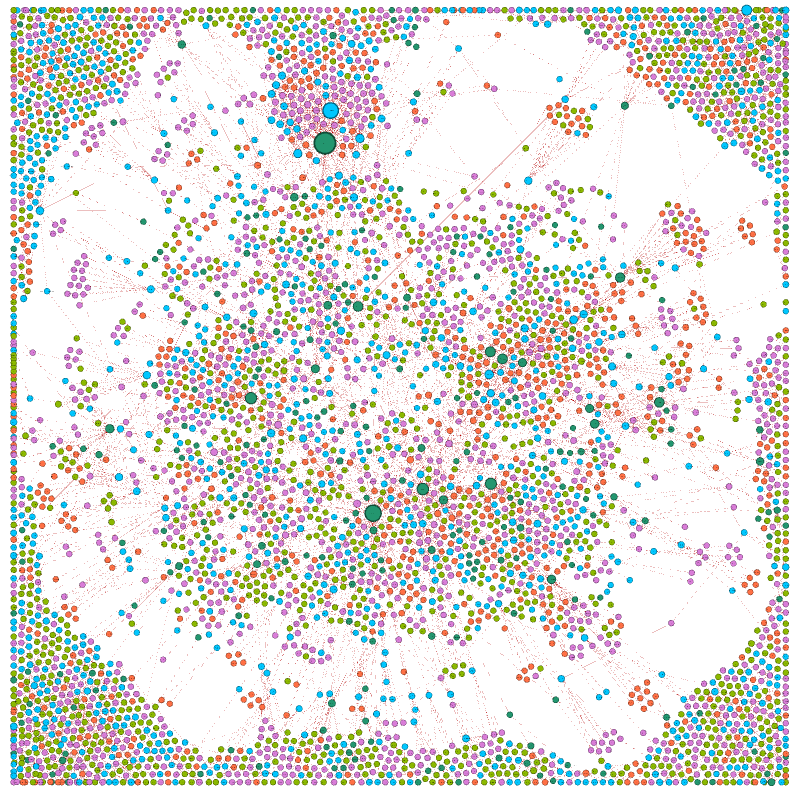
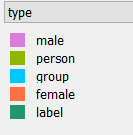
Strengths-

The graph is visually appealing and the colours are well balanced. Clear and cogent labelling allows the user to quickly identify the most prominent characters and the legends provided along with the graph allow the viewer to understand the dynamics of the story. Most of the nodes have enough space between them which is good for readability.

Weakness-

The node colour also could’ve depicted the chapter page the actant was introduced. Since this information is lacking, it’s hard to tell which actant belongs to the same timeline as the protagonist.

Some of the edge labels are overlapping which makes it hard to understand the type of interaction. This is because there are multiple edges between the same nodes representing multiple actions. This creates the need for viewing the dataset to determine the multiple actions.

**KPOP Dataset**

**Node colours**

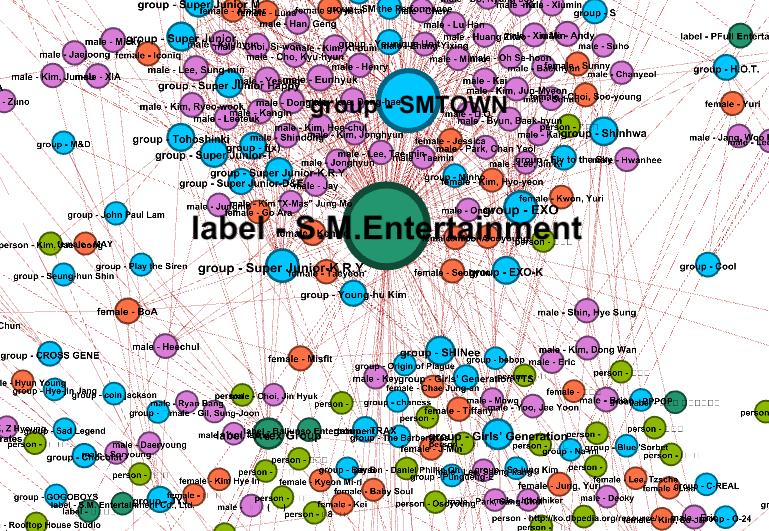
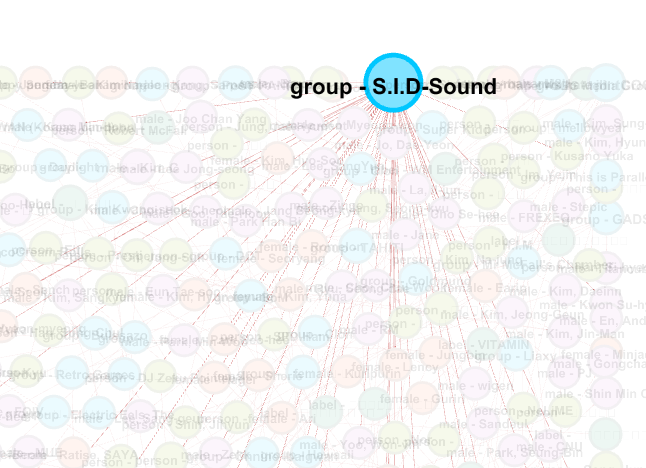
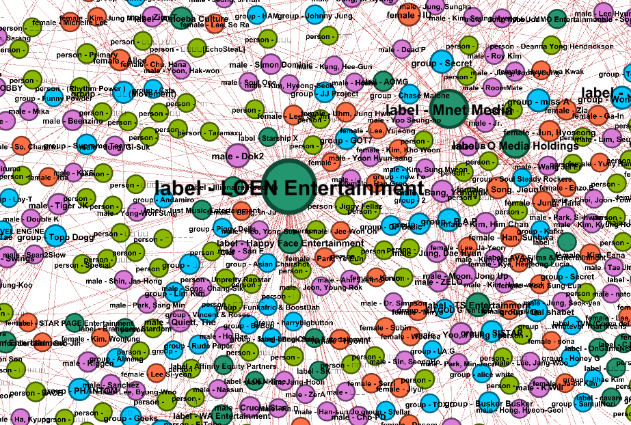
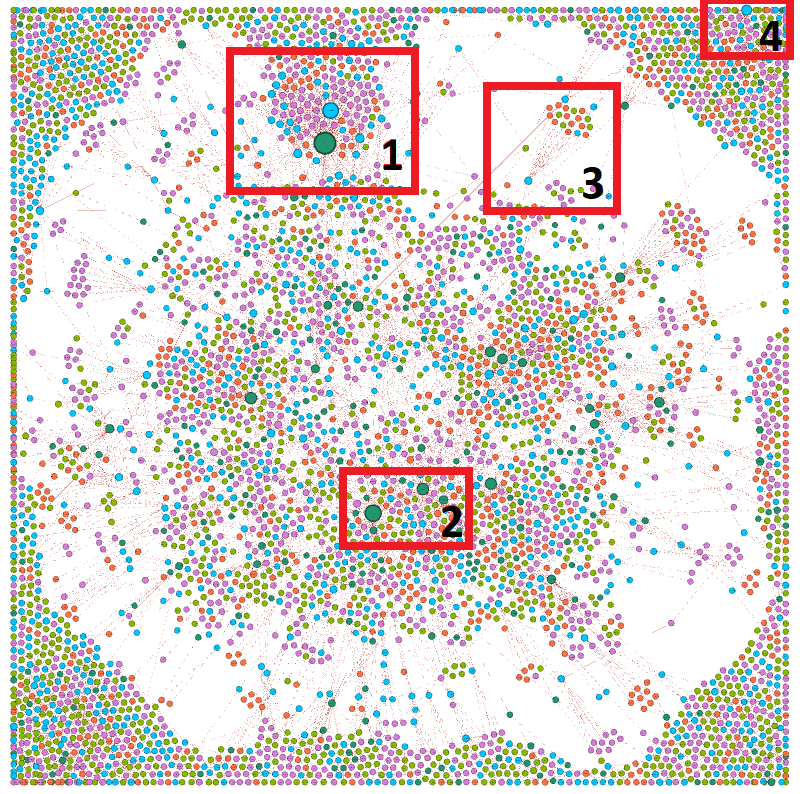
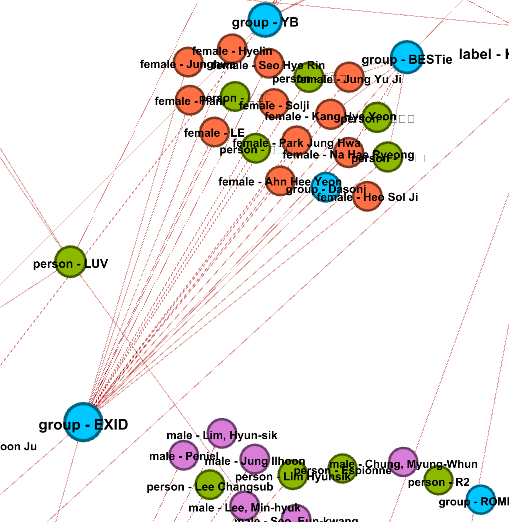
Figure 1 Figure 3

Figure 2 Figure 4

**1. Tools & layouts**

No pre-processing steps were performed. However, all nodes that did not have any edges with other nodes were filtered out from the visualization.

Gephi was used to visualize the data. Gephi was chosen for visualizations as it is intuitive to use and supports different colouring of nodes and edges.

**2. Implementation**

The nodes and edges were imported from the csv files into Gephi. The colour of node was achieved by partitioning on type (Represents Label, Person, Group, Male and Female). The edge colour gradient is based on weight of the edges. Size of the node is based on the degree of the edges. Higher number of edges on a node indicate higher number of associations with other entities.

The graph was achieved by implementing – Yifanhu Algorithm, openOrd Algorithm, Label adjust, NoOverlap stage and expansion stage to space out the nodes.

Yifanhu was the choice of graph algorithm because it generally works best with large datasets due to quick computation. Even in my experience, Yifanhu computed the visual quicker than Force Atlas. It was followed by openOrd algorithm as it helps in uncovering clusters in the graph. Label adjust, NoOverlap and expansion stages were added to make the graph more visually appealing.

**3. Visual Analysis: storytelling, sense-making**

The dataset is massive and has numerous nodes and edges. At first glance, it is prominent that the nodes which have the most relations are record labels. It is natural because record labels in general have numerous artists under contract. On drilling further down, it can be uncovered that “S.M, entertainment” label has the most number of relations. Figure 2 also shows that the group with the largest number of relations is “SMTOWN”. This could be because they are managed by “S.M, entertainment” (Largest record label). These associations can be between any type of entities- Label – label, label -artist/group, artist – group, artist-artist or group-group.

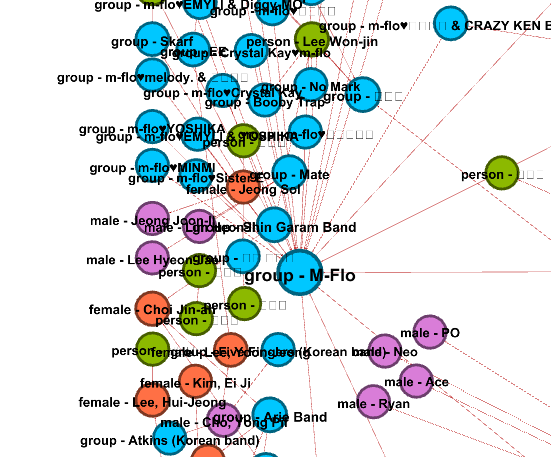
Noticeably, “Loen Entertainment” is “S.M, entertainment”’s direct competitor as it clearly has the 2nd largest number of relations(Figure 2 for reference).The artist group “EXID” has a huge female fan following and it is evident in figure 3. Figure 4 introduces us with group “S.I.D. sound” with the 2nd largest number of relations in the group category.

**4. Self-evaluation: strengths and weaknesses**

Strengths

The graph is very dense with information but simply zooming in allows the user to drill down at specific nodes to understand their dynamics. Labelling and colouring of nodes allows the viewer to quickly identify the entity type and it’s relations. The relations might be ambiguous but observing the node colous aids in the understanding of the type of relation between the nodes.

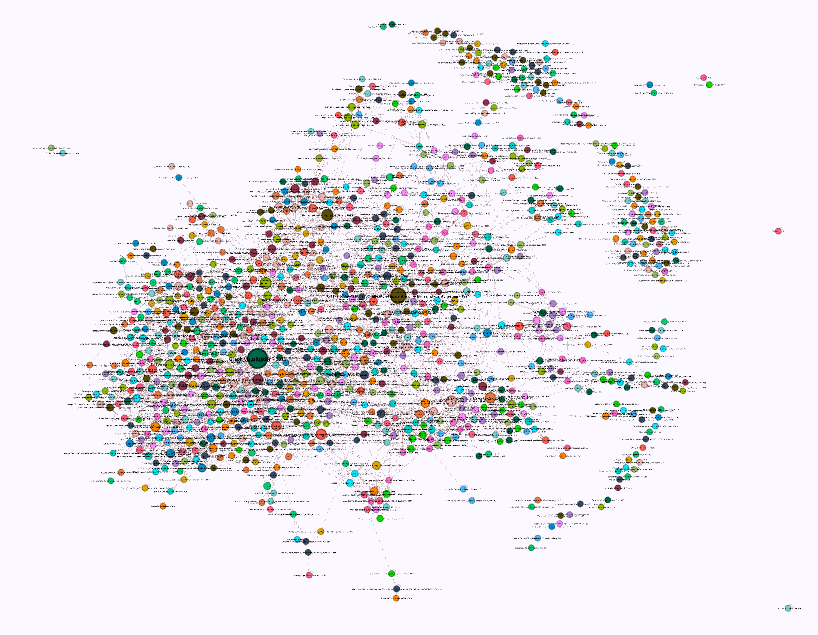
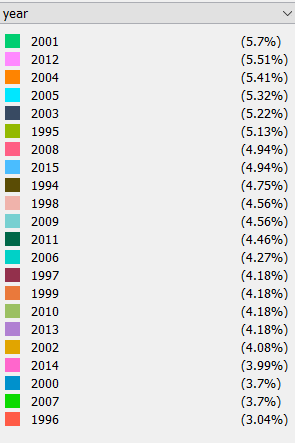
Observing the nodes neighbours gives an idea of whether the artists are engaging in more associations with other artists or have a huge fanbase. For example in the figure 5 below, it is evident that group “M-Flo” engages with other groups frequently.



Drawbacks

The graph is huge and highly complex. Viewer has to zoom into different areas and spend time to understand what the graph is about. It also seems cluttered because of the presence of extremely high number of nodes and edges.

The type of associations can be inferred by observing the neighbours of the nodes. However, the viewer can’t segment the type of associations quickly. This is not the most ideal because it may appear that an artist has a large fanbase because of higher number of persons in it’s vicinity, but it might be because they collaborate with other artists more often. These associations aren’t quickly observed because they are placed further apart.

**Type of Graph – GD-citations**

**Node colours**

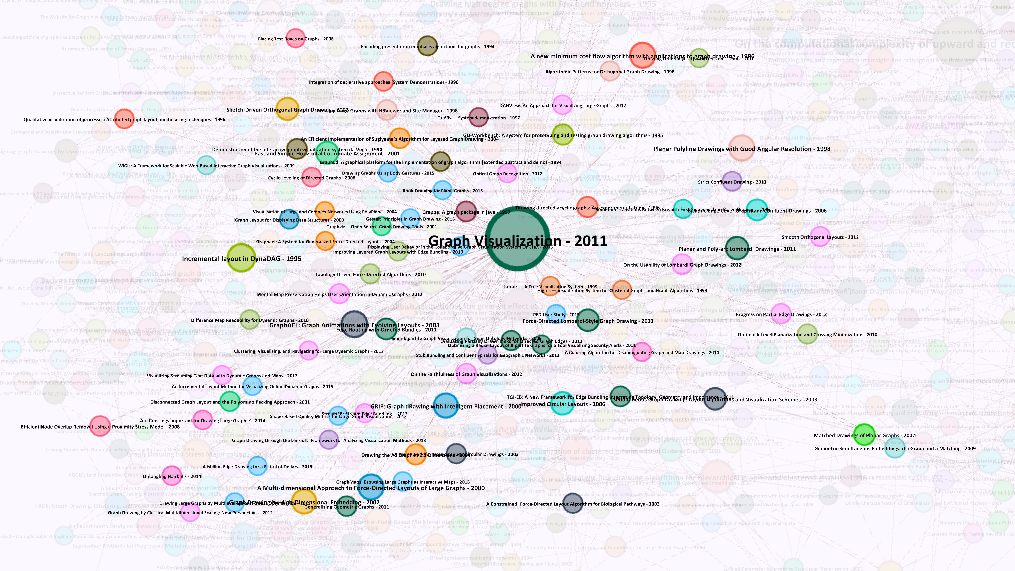
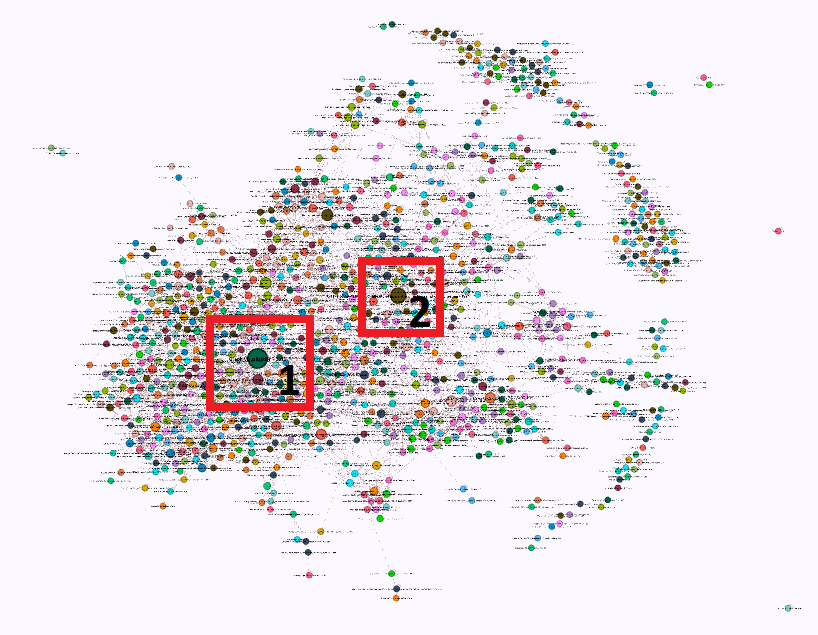


Figure 1



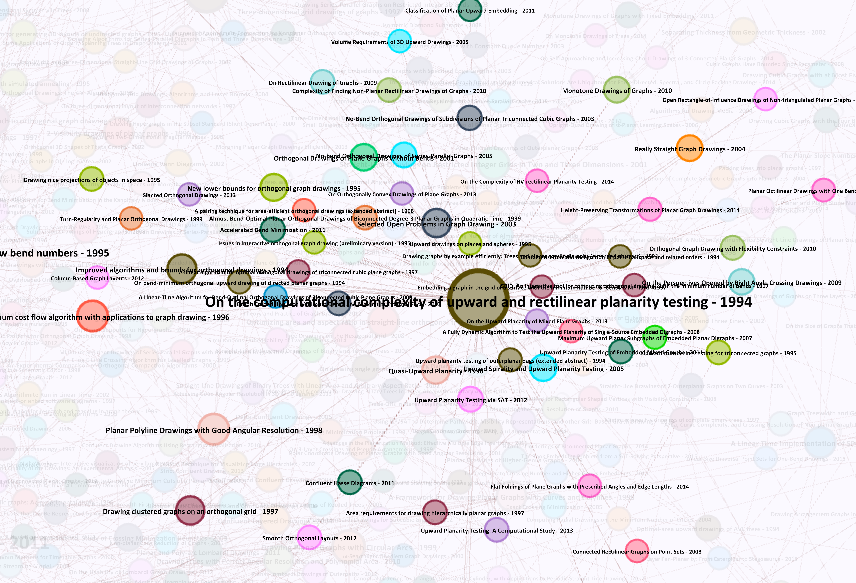


Figure 2

1. Tools & layouts

Data Pre-processing tool - Microsoft Excel.

Visualization tool – Gephi.

2. Implementation

The dataset was in the XML format. In order to import it in Gephi, it had to be converted into a csv file. After converting the dataset into csv, the ID’s available in the dataset became the nodes. However, to get the edges data, further pre-processing was required.

The “cites” and “citedBy” columns in the data were transformed into rows. Pivot in Excel was used to map research papers with their corresponding citations. The source was decided to be the main paper, and the target would be the paper that was getting cited by the main paper. In practice, if A-> B then it means that the paper B is cited by A.

After importing the pre-processed edge dataset, the graph above was obtained via the following transformations - Force atlas algorithm, expansion stage, NoOverlap stage and label adjust.

Force Atlas was chosen because the dataset is not too large and FA reliably shows the relationships in the data with fewest possible biases (Anon., n.d.). Expansion, NoOverlap and label adjust stages were added to improve the visual aspect of the graph.

Size of the nodes represent the number of citations it has received. Colour of the nodes represent the year of publication of that paper.

3. Visual Analysis: storytelling, sense-making

The graph nodes are research papers and the edges represent their citations. The node at the target of an edge is the paper that is being cited by the node at the origin. Therefore higher the number of incoming edges, higher the number of citations. From the figures 1 and 2 it is evident that papers “Graph Visualization” (in 2011) and “On the computational complexity of upward and rectilinear planarity testing” (in 1994) are the most cited papers. The legend image indicates that majority of the papers were published in 2005. Even though GV paper was the most cited, it can be observed that the 2nd most sighted paper was cited by papers that went on to receive more citations. This can be attributed to that fact that the 2nd paper was published in 1994 while the first one was published later on.

4. Self-evaluation: strengths and weaknesses

Strengths

The graph only displays Title and year of the paper published, and it’s neighbours indicate the type of papers that cited the paper as well as the frequency of citation of those papers themselves. The papers with 0 citations are not filtered out and given due recognition.

Weakness

The dataset is not overly complex but only conveys information about the year and number of citations. The nodes can be partitioned according to the institutions they were published at, but that would be difficult to portray because a node cannot have multiple colours. Moreover, the data only displays the year the paper was published but does not allow the user to determine the years it was cited after. Due to this the viewer is unable to gauge the current relevance of the publication.

# References

Anon., n.d. *Gephi Tutorials.* [Online]   
Available at: https://gephi.org/tutorials/gephi-tutorial-layouts.pdf