**Visualisation of relational data with different layout techniques supported by Cytoscape**

**Introduction**

Network analysis and visualization is an important technique for allowing researchers to interprettheir data from a different perspective. In the field of graph drawing, a good visualization emphasizes readability and promotes understanding (Bennett et al. 2007).

This report focuses on visualizing datasets which can be visualized as networks or graphs. Graphs are abstract structures that are used to model relational information. A graph consists of a set of vertices (V) and a set of edges (E) that connect those vertices together.. A graph can have directed or in directed edges also edges can have weights or other properties. Networks can be of type social networks or general networks. Dedicated algorithms, called layouts, calculate the node positions and display the data on 2D/3D. We will discuss some of those algorithms in this report.

To make the graphs visually pleasing certain Aesthetic criteria are followed such as:

Node metrics: Clustering of similar nodes, Distribution of nodes evenly, Maximize node orthogonally.

Edge metrics: Minimize edge crossing, minimize edge bends, Maximize edge orthogonality.

Overall Layout Metric: Maximize convex faces, maximize consistent ﬂow di-rection, keep correct aspect ratio, Minimize area (Bennett, Chris & Ryall, Jody & Spalteholz, Leo & Gooch, Amy 2007).

**Cytoscape**

Cytoscape is a Java based platform which is open source used for visualizing graphs, networks and biological interactions of molecules. Although originally intended for biological science, Cytoscape is now a general platform for complex analyses and network visualization. Cytoscape 's core framework includes a simple set of functions for data integration, visualization and analysis. (manual.cytoscape.org)

Users communicate with the image to control the revealing structures, shapes and colours.The goal is to help analysts make observations, discover patterns intuitively and during data collection identify singularities or defects in the structure.Cytoscape offers simple features for incorporating arbitrary data into the database, a visual representation of the database and merged data, sorting and filtering tools.

Cytoscape allows users to load data in various formats or public datasets like NDEx, PSICQUIC, STITCH and Pathway Commons. Network then can be filtered by manually selecting nodes or applying a filter based on node or edge attributes. Layout of the network can be altered using various Layout algorithms supported by cytoscape, other algorithms can be applied by using cytoscape apps.

The network can be made visually pleasing by using “Style” Tab of Cytoscape. Where colours, borders, transparency and many more properties can be altered. Finally, legends can be created for better understanding for the network.

Accepted file formats in Cytoscope:

* Simple interaction file (SIF / .sif format)
* [Cytoscape CX](https://github.com/CyComponent/CyWiki)
* SBML
* BioPAX
* Nested network format (NNF / .nnf format)
* Graph Markup Language (GML / .gml format)
* XGMML (extensible graph markup and modelling language).
* GraphML
* [Cytoscape.js JSON](http://cytoscape.github.io/cytoscape.js/#notation/elements-json)
* Delimited text
* Excel Workbook (.xls, .xlsx)

**Data**

For this report 3 datasets in GML format have been used to visualize using different techniques, namely:

**Political Books Network (polbooks.gml)**

A network of books on American politics published at the time of the 2004 presidential election and distributed exclusively through Amazon.com. Edges reflect repeated co-purchasing of the same as indicated by the "customers who bought this book also bought these other books" feature on Amazon.

Nodes were provided with values "l", "n", or "c" to indicate whether they are "liberal", "neutral", or "conservative". Those alignments were allocated separately by Mark Newman based on a reading of the descriptions and reviews of the books found on Amazon.

V. Krebs assembled the network and it is unpublished (Ref <http://www.orgnet.com/> )

The dataset consists of **105 nodes and 441 edges**.

**Political Blogs Network (polblogs.gml):**

A direct network of hyperlinks about US politics between weblogs, documented by Adamic and Glance in 2005. Please quote L. A. Adamic, n. Glance, "The Political Blogosphere and the 2004 US Election," in WWW2005 Conference on the Weblogging Ecosystem (2005) Proceedings

Node "value" attributes indicate political leaning according to:

0 (left or liberal)

1 (right or conservative)

Data consists of **1490 nodes and 19090 edges.**

**Power Grid (power.gml):**

A network that represents the topology of the US Western States Power Grid. Compiled data by D. Watts y S. Strogatz and made available. Please quote D. J. Watts y S. H. Strogatz (1998), Nature 393, 440-442.

Data consists of **4941 nodes and 6594 edges.**

**These datasets are of GML format, hence, can be directly imported in Cytoscape using File -> Import -> Network from file.**

**Visualisation and Analysis**

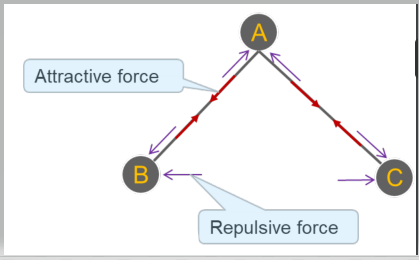
Different visualisation methods can be applied to datasets using “Layout” Tab in Cytoscape. These visualisation methods are based on popular algorithms used for Tree or network visualizations along many years. This report discusses some of these “Layouts” present in Cytoscape.

**Forces Directed Methods - Prefuse Force Directed Layout**

Also known as spring embedders, these algorithms measure a graph 's layout using only information associated in the graph's own structure, instead of relying on domain-specific knowledge. *(Kobourov, Stephen. 2013)*. Graphs drawn using these algorithms tend to be aesthetically pleasing, display symmetries and help to create cross-sectional layouts for planar graphs. *(Kobourov, Stephen. 2013).*

Network nodes are perceived as particles repelling each other, for example electrons. The internode **connections are handled like metal springs** connected to the node pair. Such springs repel their end points or attract them according to a function of force. **The layout algorithm sets node locations in such a way that the sum of forces in the network is minimised.**

This layout is based on the “force-directed” paradigm aka spring-embedder or energy-based placement algorithm. The algorithm was introduced as part of Jeff Heer's prefuse toolkit. It is a fast algorithm and can provide a very visually pleasing layout with the right parameters.

****

Type of forces acknowledged in this algorithm: -

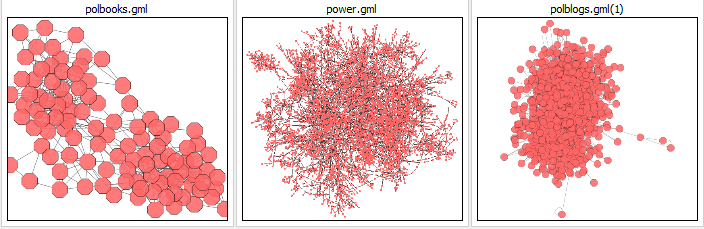
**Attractive:** It is through each of the edges and it is proportional to the shortest paths.

**Repulsive**: It is between all set of 2 vertices and it exists only between close vertices.

**Other:** Centre of gravity, magnetic field, angular, from the boundary.

**Advantages**: Good for implementation with general graphs, aesthetically pleasing, fast and simple to implement, Empirically strong symmetric and structural representation, good results for small and medium-sized graphs (50 to 500 nodes)

**Disadvantages**: Non- predictable layout, slow with high cost in computation, too much spinning to get full shape, Not well scale on larger graphs.



*Figure 1: Prefuse Force Directed Layout of 3 datasets with 105,4941,1095 nodes and 441,6594,18962 edges respectively.*

**Style used**: - PSIMI 25 with black borders.

**Filter used**: - Degree filter (In and out) between 2 and 594 inclusive for polblogs.gml dataset.

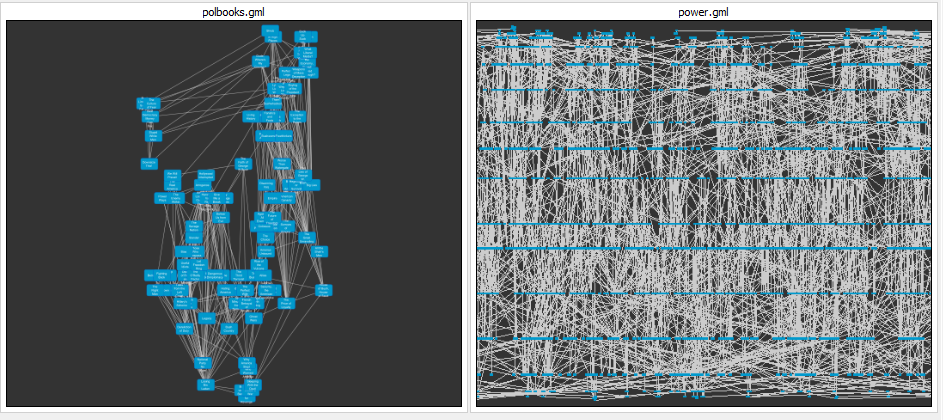
**Hierarchical Drawing Algorithm - Hierarchical Layout**

Hierarchical graph drawings are useful for providing insight into hierarchical structures in complex systems *(Nikolov, Nikola 2016*)

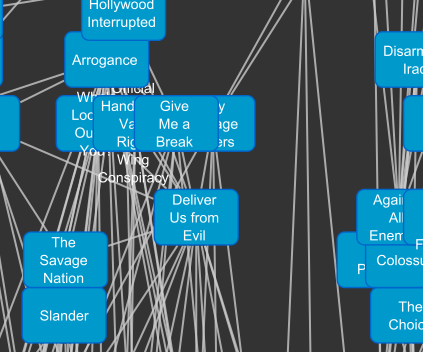
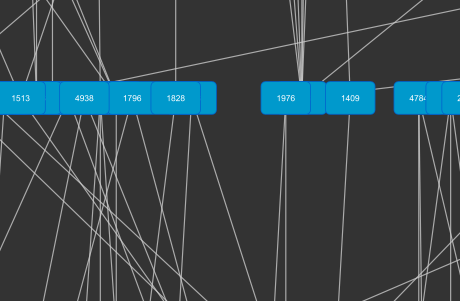
The algorithm of hierarchical layout is perfect for describing the main path or the **"flow" within a network.** Nodes are arranged in hierarchically ordered layers, and the node order is chosen within each layer to minimize the number of edge crossings. Often used while depicting flow of information, workflows and process diagrams.

**Advantages**: Many edges flow in the same direction, number of crossings is minimised, Display very simply the movement of information and Hierarchy

**Disadvantages**: Not ideal for general graph and not suited for large graphs as it doesn’t scale well.



*Figure 2: Hierarchical Layout of 2 datasets with 105,4941 nodes and 441 ,6594 edges respectively.*

*Polbooks.gml power.gml*

*Figure 3: Zoom view of Hierarchical Layout*

**Style used**: -Default layout with rectangular nodes with grey background and white edges.

**Filter used**: - Degree filter (In and out) between 2 and 594 inclusive for polblogs.gml dataset.

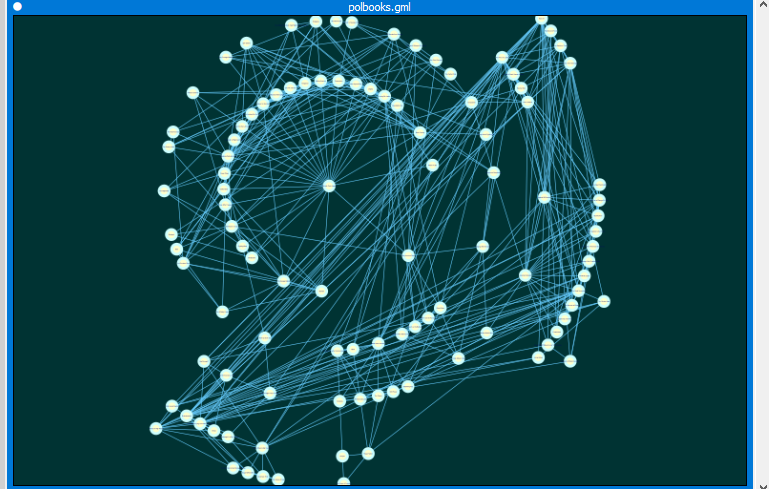
**Radial View Algorithm – yFiles Radial Layout**

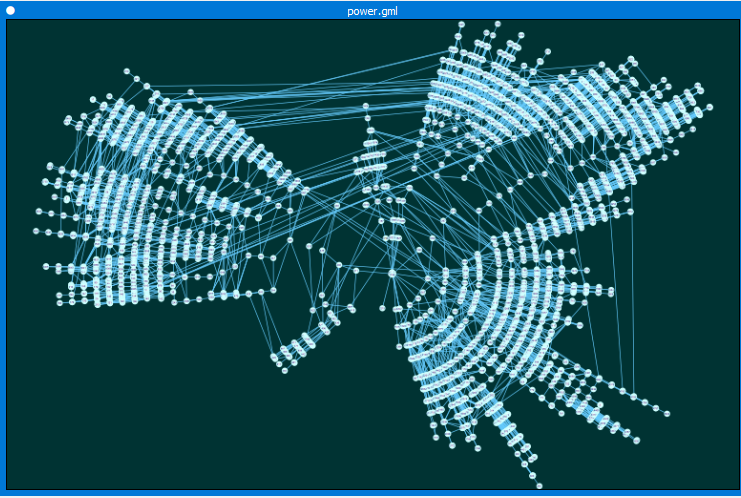
Radial View is a **connection approach** to visualize trees.

The Radial algorithm **places nodes around a specific core in virtual concentric circles**. This algorithm illustrates tree structures in a diagram. It includes placing children of a subtree recurrently in ringed wedges. The angles of such wedges are proportional to the number of sub-tree leaves. To stop node angle to be very large Radial View Algorithm forces all wedges to remain convex. It helps to display a tree like structure which expands radially.

**Advantages**: It is a simple algorithm, which behaves very good, the layout is predictable, aesthetically pleasing, direct relationship of nodes are visible.

**Disadvantages**: Complex geomantic forms due to which interactive features are more difﬁcult to develop and to apply (Burch, Michael & Weiskopf, D. 2014), doesn’t utilise space optimally.





polbooks.gml power.gml

*Figure 4: Radial Layout of 2 datasets with 105,4941 nodes and 441,6594, edges respectively.*



*Figure 5: Zoom view of Radial Layout for polbooks.gml and power.gml respectively.*

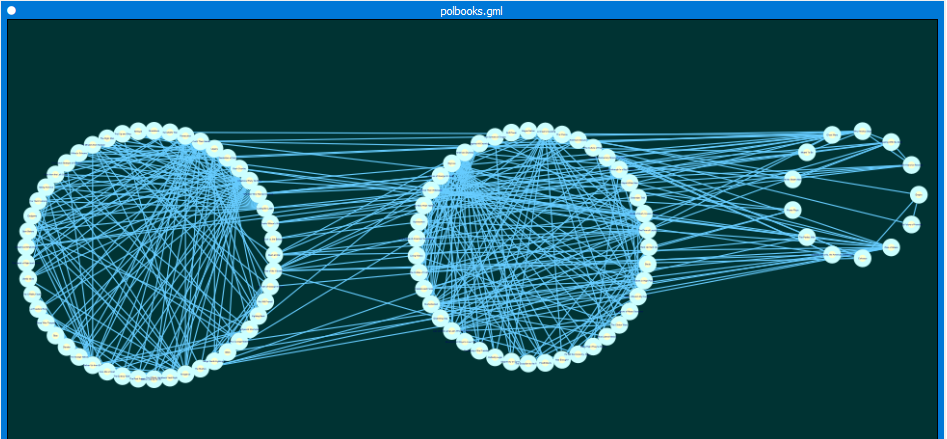
**Style used**: - Ripple

As observed Prefuse **f**orced directed, Hierarchical and radical Layouts are optimal for our dataset polbooks.gml having 105 nodes and 441 edges **but not optimal for nodes above 1000 in our dataset power.gml and polblogs.gml** as they are being visualized as “**hairballs**”. The edges are **overlapping**, and the space is too crowded to analyse the network.

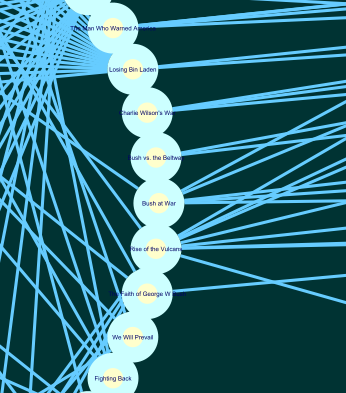
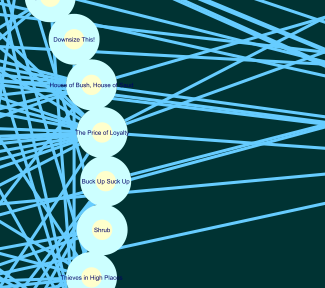
**For these datasets we use aggregation and abstraction techniques and to show a higher level of tropology**.

**Group Attributes Layout**

This layout places all the nodes having the same value of column around a circle in the network. (manual.cytoscape.org). In this layout all nodes with the same value for a specific column are placed together around a circle(manual.cytoscape.org).



*Figure 6: Group attribute layout for polbooks.gml having 105 nodes and 441 edges grouped by value column*



*Figure 7: Zoom view of Group attribute layout for polbooks.gml*

Here we have used group attribute layout to group by value of node which can be l, n or c as described in Data Set section of the report. We can see three groups clustered together.

Here we can observe that by grouping the nodes together i.e. aggregation we have clusters of nodes which are more visually pleasing for big graphs.

|  |  |
| --- | --- |
| *Figure 8: Group attribute layout for grid.gml having 4941nodes* *and 6594 edges.* | *Figure 9: Zoom view of Group attribute layout for grid.gml* |

For grid.gml grouping has been done with node degree i.e. the no. of vertices adjacent to a vertex V. We can see nodes clustered together with degree 2,3 so on.

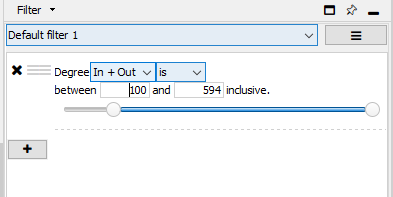
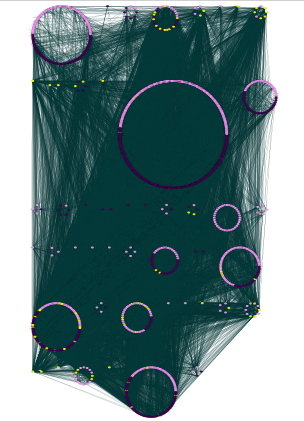
|  |  |
| --- | --- |
| *Figure 10: Group attribute Layout for polblogs.gml with. 1490 nodes and 19090 edges grouped by source column* | *Figure 11: Zzoomed image of Group attribute Layout for polblogs.gml* |

For dataset polblogs.gml Group Attribute Layout is used with grouping based on “source” column. The nodes are clustered together based on their source. Due to structural clustering this

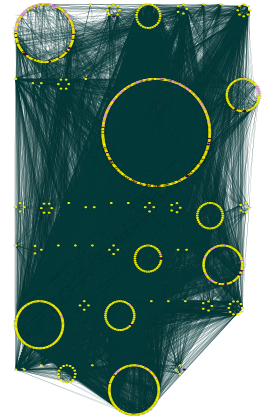
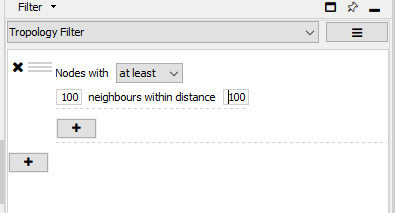
visualization is offering better analysis.

**Other Techniques for better visualization of big graphs**

1.Filtering nodes based on attributes can help to create a sub network for better analysis. We can **filter based on column data** and other attributes. Degree filter can be applied to nodes to match degree of nodes. In-filter and out-filter options can be selected.

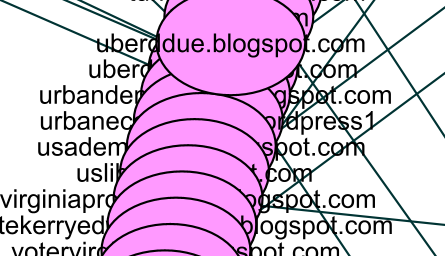
  
*Figure 12: polblogs.gml (1490 nodes and 19090 edges) with degree filter*

2. Tropology Filters may be used to align nodes with a certain number of neighbors within a specified radius. Showing a **high-level view of tropology** and help to better visualize big graphs.



*Figure 11: polblogs.gml (1490 nodes and 19090 edges) with tropology filter*

3.Extremely dense networks can be better visualised by **controlling the opacity of vertices**.



*Figure 12: polblogs.gml with transparency 70 and 255(none) respectively.*

**Discussions**

**Visualisation Tools**

There are many other tools online available to visualize networks and graphs Cytoscape is one of them.

**Cytoscape Cons as observed**

Through interaction with Cytoscape it is very prominent that it is mostly used for biological research and biological network visualization. While there are many public networks to search from built in search in Cytoscape for biological networks. Other networks must be manually imported through accepted formats or table files such as csv or xls. Large and complex graphs take too much computation time.

**Cytoscape Pros as observed**

The interaction with network is very smooth, applying styles and selecting nodes and edges is very simple. Many Layouts are available to apply, and other layouts can be applied by using Cytoscape Apps.

Discussion about some other tools available for network visualization:

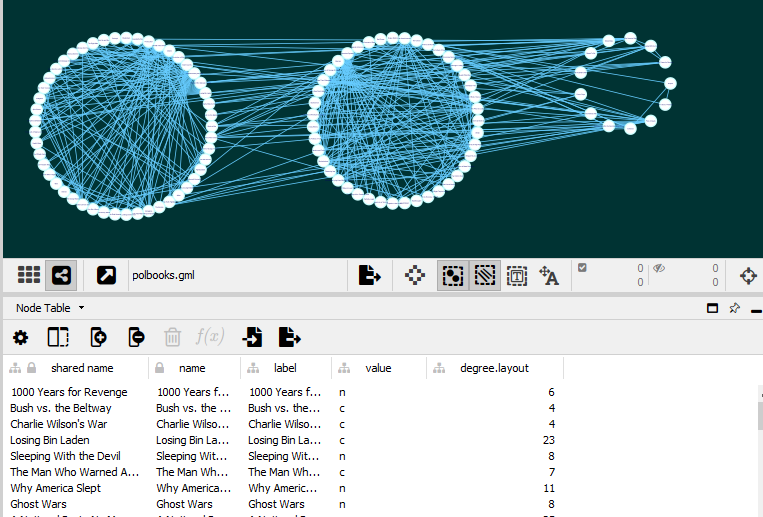
| **Tool** | **Gephi** | **Kumu** | **Cytoscape** | **igraph** |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
| **About** | An open-source program allowing the visualization of large networks to be visualized and analysed. Designed with the concept of being the Network Visualization Photoshop, its main strength is to represent networks but there are also simple statistical properties. Uses a 3D modelling engine, which shows broad real-time graphs. Seen often in social science-related network research and cultural studies. | A web-based application for the analysis and monitoring of relationships. Developed to make people appreciate more quickly complex partnerships. | An open source Software framework built to imagine networks of molecular interactions. Though initially designed for biological research, and increasingly used in biology and health sciences, it is a general tool for complex network analysis and visualization, providing a basic set of data aggregation, analysis and visualization functionalities. | An open-source repository of network research software with focus on performance, portability and user-friendliness. |
| **Platform** | Java | Web (cloud) | Java | R / Python / C/C++ |
| **Pricing** | Free | Free for public projects | Free | Free |
| **Tractable number of nodes** | Networks up to 100,000 nodes and 1,000,000 edges | Works best up to 10,000 elements and connections | Depends on memory size of machine | More than 1.9 million relations (without attributes) |
| **Pros** | - No programing knowledge required | - No programing knowledge required | - No programing knowledge required | - Handles very large datasets |
|  | - Handles relatively large networks (depending on user environments) | - Very easy to use, very simple UI. Great for students that lack in-depth awareness of network research and visualization | Provides a massive collection of additional plugins (> 250) expressly developed to resolve and automate biological research. Best for biological network analysis | - Powerful analysis functions |
|  | - User friendly and intuitive UI |  | - User friendly UI | - Advantages in using R (cross-platform, import/export any form of data, has a huge library of packages) |
|  | - Produces the highest quality and most attractive visualizations |  | - Produces high quality visualizations | - Can be used together with other SNA libraries such as Statnet |
| **Cons** | - Not strong network analysis ability | - Not strong network analysis ability | - Requires a steep learning curve to use it and plugins for more advanced tasks | - Requires some knowledge in R programming |
|  |  | - Cannot handle large networks |  | - Text-based, not user-friendly at all |
|  |  |  |  | - Not a visualization tool. Hard to make attractive visualizations |

The key obstacles to the usage of any of these simulation techniques are the scale of the network to be examined, as well as the amount of information and expertise required to allow effective use of the software.

**Graph Analysis**

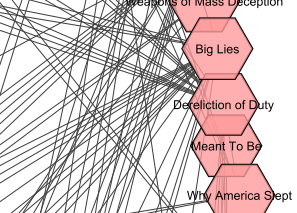
**1.Dataset polbooks.gml**

From Group Attribute Layout we can observe that biggest cluster of nodes with have value “c” i.e. are of the “conservative” group. This implies that **the most sold books** on amazon from this dataset are of **“conservative” political philosophy**, followed by liberal and neutral. Neutral has only 14 nodes out of total 105 nodes of dataset i.e. the **least books sold were of “neutral” political philosophy**.



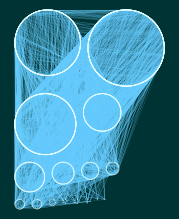
*Figure 13: clusters of nodes having value “c”, “l” and “n” respectively*

Also, we can observe that the graph edges are most dense in “liberal” cluster that implies **frequent co-purchasing of books of liberal philosophy by same buyers.**

We can also observe that “Big Lies” and “Dereliction of Duty” are some of nodes having most edges that implies that these are popular books of which many copies were sold.  
*Figure 16: most connected nodes of polbooks.gml*

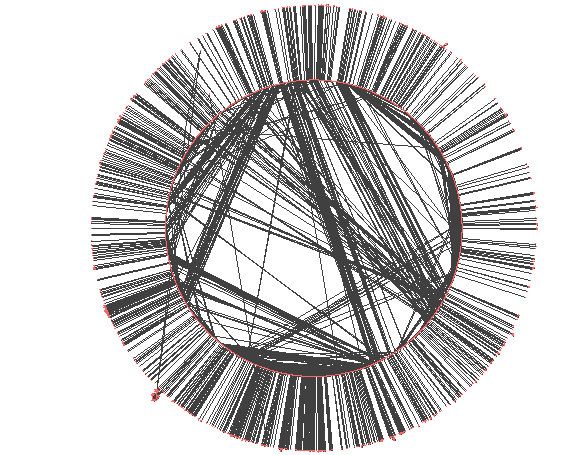
**2.Dataset grid.gml**

We can see nodes having degree 1 to 19. The biggest clusters having degrees 1 and 2, we can say that the network of grids is moderately connected.



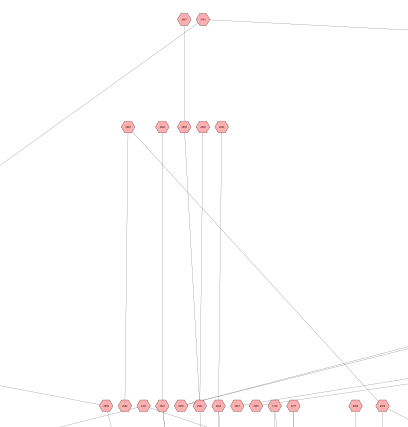
*Figure 17: Clusters of nodes having degree 1,2 and so on.*

Through hierarchal and circular layout, we can differentiate between hubs and end nodes in this tropology of power grids.



The inner circle denotes nodes which are central can also be hubs and the outer circle consists of end nodes of tropology.

*Figure 18: circular layout for power.gml*



Through Hierarchical Layout we can observe the central nodes or hubs which are in top layers and end nodes which are in bottom layers

*Figure 19 Zoom view of Hierarchical layout of grid.gml*

**2.Dataset polblogs.gml**

From grouping this dataset with column “source” it is observed that is the biggest source of blogs is Blogarama followed by BlogCatalog and eTalkingHead.

Grouping this dataset with column “value” which could be of type 0 (left or liberal)

or 1 (right or conservative) it is observed that the no. of blogs of each political philosophy are almost equal. The network is very dense as it has 19090 edges which is due to the suggestion similar blogs by these source websites.

|  |  |  |
| --- | --- | --- |
| *Figure 20: Group attribute layout for value 0 or 1* | *Figure 21: Group attribute layout on column “source”for polblogs.gml* | *Figure 22: Circular Layout for polblogs.gml* |

**Conclusion**

The result indicated that small, medium and large datasets can be visualised by using Cytoscape through applying required techniques. Cytoscape is user-friendly, easy to understand and use and helps to develop visualizations by people from non-programming backgrounds as well. In the end, visualizations should save time on data discovery and should not hinder data comprehension (Gough 2014). As careful planning, arrangement and screening methods must be used to finish with a satisfactory outcome, Cytoscape also offers to apply different layout, filters and styles to provide visualizations that are easy to analyse.

References:

Bennett, C, Ryall, J, Spalteholz, L & Gooch, A 2007, 'The aesthetics of graph visualization', Computational Aesthetics, vol. 2007, pp. 57-64.

Kobourov, Stephen. (2013) “Force-Directed Algorithms”.

Nikolov, Nikola. (2016). Sugiyama Algorithm. 10.1007/978-1-4939-2864-4\_649.

Burch, Michael & Weiskopf, D. (2014). On the benefits and drawbacks of radial diagrams. 10.1007/978-1-4614-7485-2\_17.

Lada A. Adamic and Natalie Glance, "The political blogosphere and the 2004 US Election", in Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem (2005).

D. J. Watts and S. H. Strogatz, "Collective dynamics of `small-world'

networks", Nature 393, 440-442 (1998).

Bennett, Chris & Ryall, Jody & Spalteholz, Leo & Gooch, Amy. (2007). The Aesthetics of Graph Visualization. Proceedings of Computational Aesthetics in Graphics, Visualization, and Imaging. 57-64. 10.2312/COMPAESTH/COMPAESTH07/057-064.

Gough, P, Ho, X, Dunn, K & Bednarz, T 2014, 'Art and chartjunk: a guide for NEUVis', in Proceedings of the 7th International Symposium on Visual Information Communication and Interaction, p. 171.

manual.cytoscape.org (2020, 07 28): <http://manual.cytoscape.org/en/stable/Introduction.html>

datasets (polblogs.gml, polbooks.gml, power.gml (2020, 07 28) : <http://wwwpersonal.umich.edu/~mejn/netdata/>

Visualization tools (2020, 07 28):<https://gwu-libraries.github.io/sfm-ui/posts/2017-09-08-sna>

About Cytoscape (2020, 07 28):<https://genome.cshlp.org/content/13/11/2498.long>