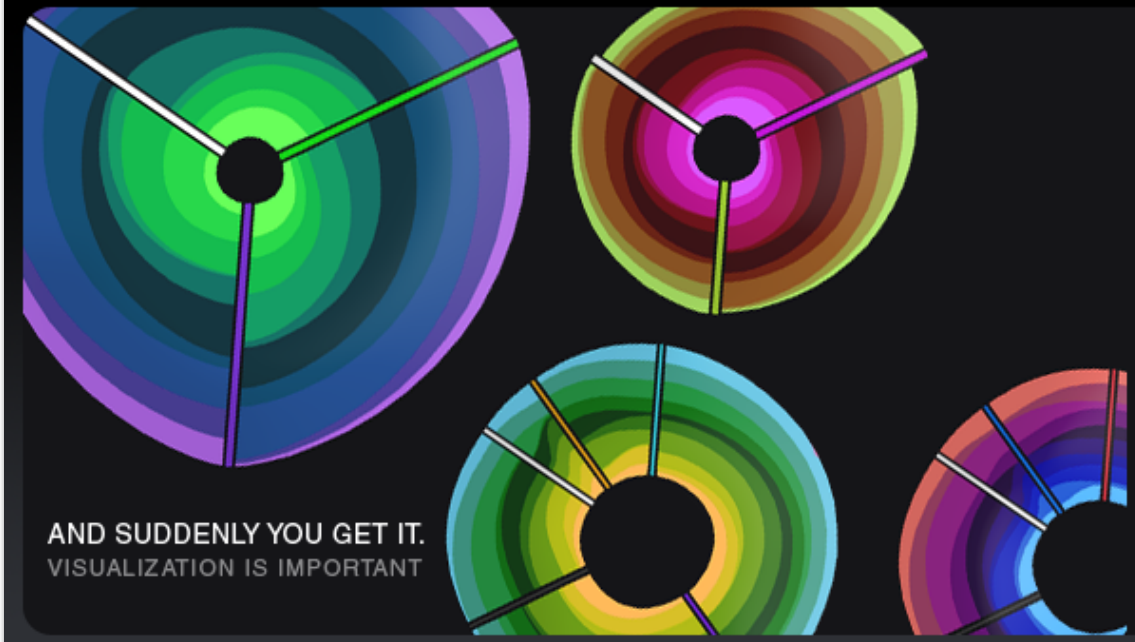


AND SUDDENLY YOU GET IT.

VISUALIZATION IS IMPORTANT

LINEAR LAYOUT FOR VISUALIZATION OF NETWORKS
THE END OF HAIRBALLS

MARTIN KRZYWINSKI

BC CANCER RESEARCH CENTER
BC CANCER AGENCY
VANCOUVER BC CANADA

Talks and software are available at http://mkweb.bcgsc.ca/linnet.

Originally presented at Genome Informatics 2010, Hinxton UK (September 17).

BC Cancer Research Center
Genome Sciences Center

A HAIRBALL

A hairball of the wooly mammoth.

/ used as the DNA source for sequencing of the mammoth

Regenerating a Mammoth for 10 Million, NYT 2008

/ visual examination of the hairball does not reveal characteristics about the mammoth

// probably hairy

// cold climate dweller

Miller W, Drautz DI, Ratan A, Pusey B, Qi J, Lesk AM, Tomsho LP, Packard MD, Zhao F, Sher A et al. 2008. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* **456**(7220): 387-390.

A BETTER VISUALIZATION OF THE HAIRBALL

This is an excellent visualization of the mammoth.

/ it communicates
// size
// shape
// environment
// prominent characteristics

/ the aggressive posture suggests the animal is not suitable for immediate domestication.

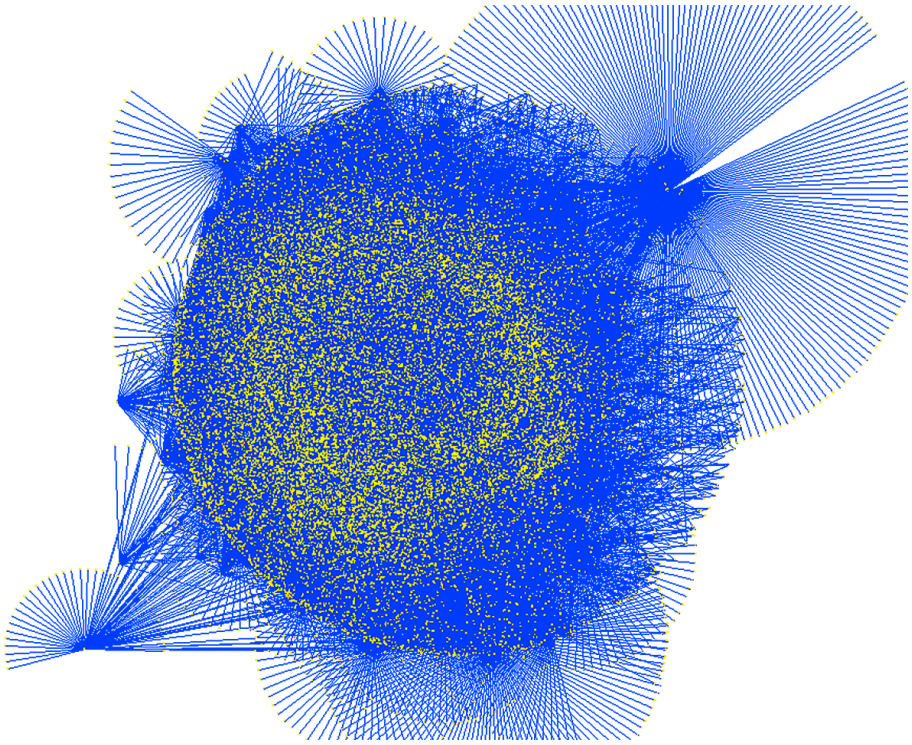
http://wild-facts.com/2010/09/22/wild-fact-718-the-snow-plow-wooly-mammoth/

A HAIRBALL

A hairball visualization of the network.

/ it tells us as much about the network, as about the mammoth – not much at all.

THIS TALK IS ABOUT GENERATING A BETTER VISUALIZATION THAN THE NETWORK HAIRBALL

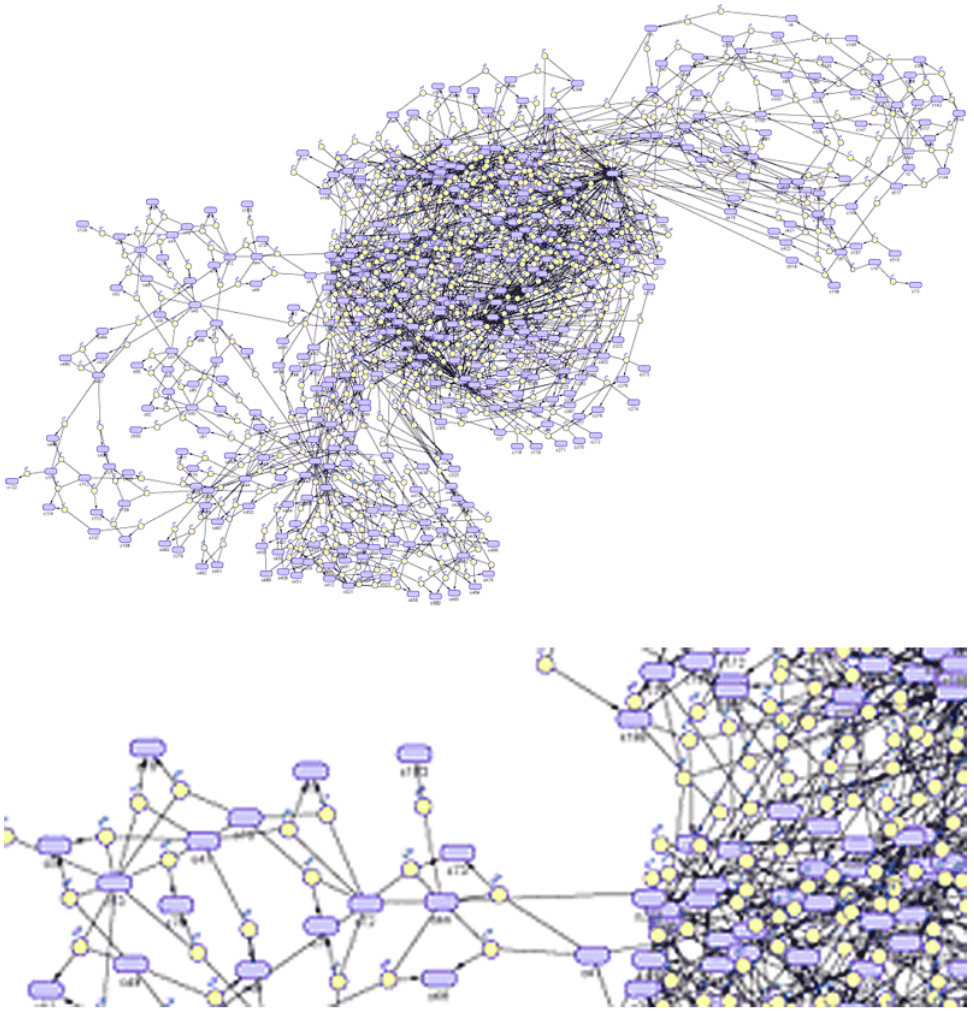


A depiction of the hyperlink network in gene wiki, rendered with Cytoscape (i9606.blogspot.com).

THE PROBLEM

conventional network visualization is unsuitable for visual analytics of large networks

so-called *hairballs*, these visualizations have significant disadvantages



A small section of a biological system. Each system comprises a convoluted network of smaller pathways, each with its own level of complexity. Inset at bottom is a magnified portion of the network, illustrating complexity.
www.mathworks.com

CONVENTIONAL NETWORK VISUALIZATION – HAIRBALLS

IMPENETRABLE COMPLEXITY

rapidly grow in visual complexity

become visually impenetrable

DEPICTIONS OF LARGE NETWORKS EXCEED RESOLUTION OF OUTPUT AND VISUAL PERCEPTION

DATA SUBORDINATE TO LAYOUT

hairball's form is determined by the layout algorithm

node and edge metadata are subordinate to layout

important characteristics of the network cease to drive the visualization and cannot be evaluated

HAIRBALL VISUALIZATIONS DO NOT CLEARLY REFLECT ASPECTS OF INTEREST

COMPARISON IMPOSSIBLE

layout algorithm is a major influence of the final visualization

similar networks may have different layouts

DIFFERENCES BETWEEN TWO HAIRBALLS DO NOT NECESSARILY REFLECT A DIFFERENCE IN THE DATA, NOR CLEARLY CAPTURE THE EXTENT OF THE DIFFERENCE

VISUAL CONFUSION

disambiguating layout algorithm from data is impossible

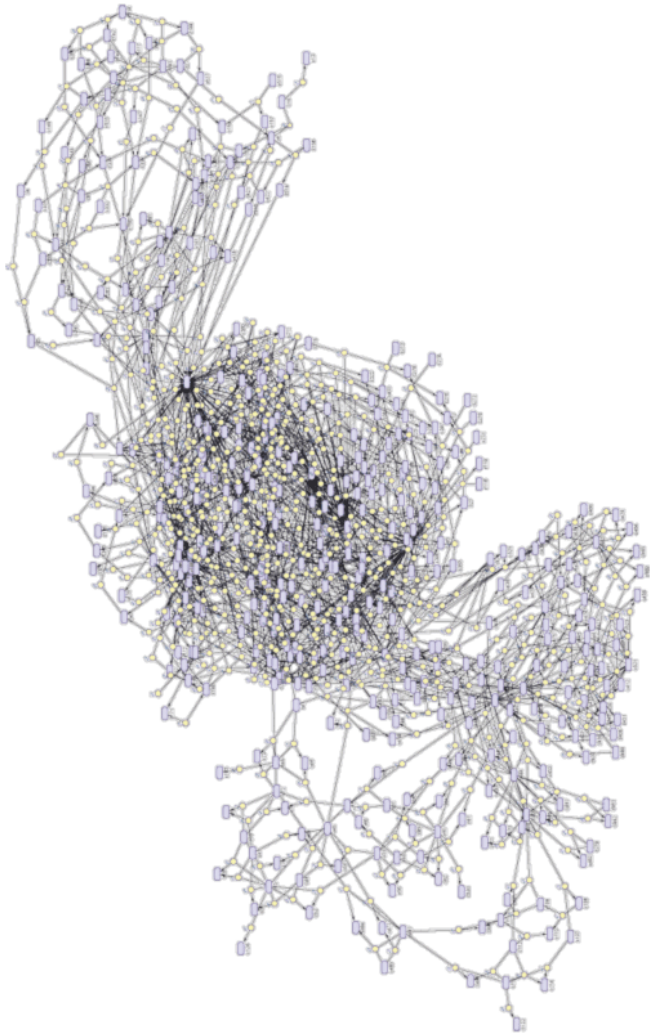
/ the hairball on the left appears to have a central hive of connected nodes

/ is this a property of the data or the layout?

/ would a different layout algorithm reveal more?

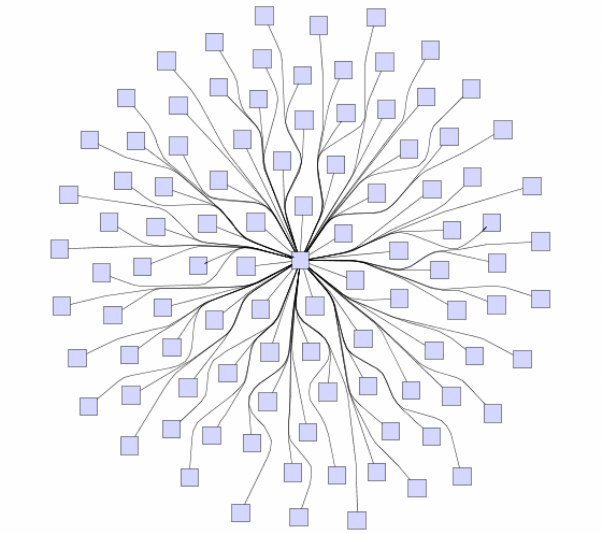
/ what is the ideal layout algorithm?

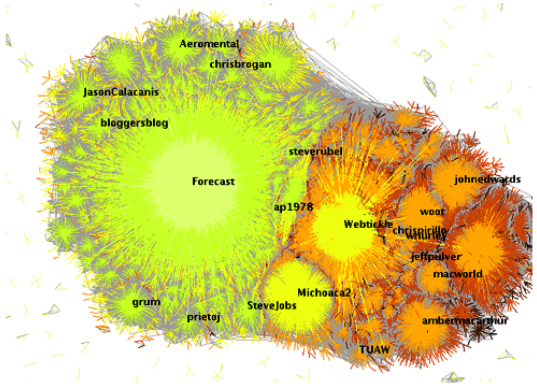
HUMAN PATTERN MATCHING ABILITIES CANNOT BE EFFECTIVELY ENGAGED

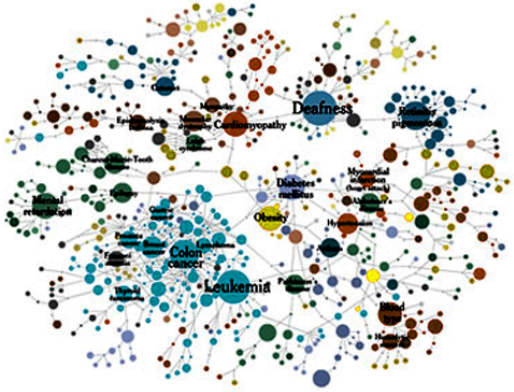


A small section of a biological system. Each system comprises a convoluted network of smaller pathways, each with its own level of complexity.
www.mathworks.com

CAPTURING IMAGINATION







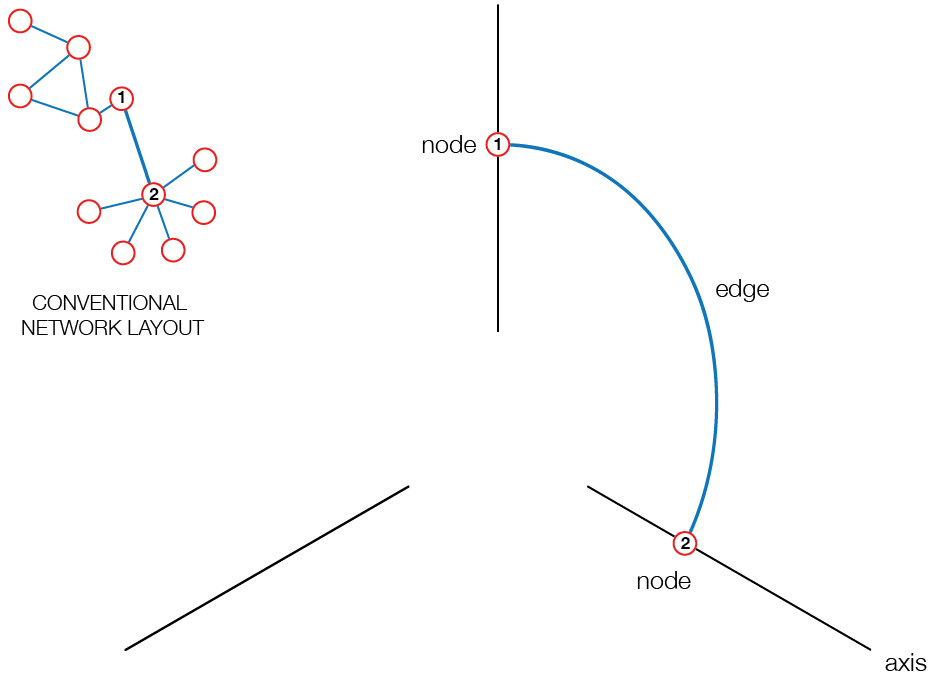
There are many algorithms to draw networks and many beautiful renditions of large networks are available. To make these layouts informative is incredibly challenging – requiring parameter tweaking or manual adjustment specific to the data set. y.layout.router Class OrganicEdgeRouter / Large Graph Layout (LGL) / Today by Cada / Mapping the Human Disseaseome (Bloch/Corum NYT 2009)

THE SOLUTION – CONCEPT

the linear network layout addresses the shortcomings of the conventional layout

/ nodes are constrained to linear axes

/ edges are drawn as curves between nodes



In the linear network layout, nodes are constrained to linear axes. Edges are drawn as curves between connected nodes.

THE SOLUTION – MAPPING

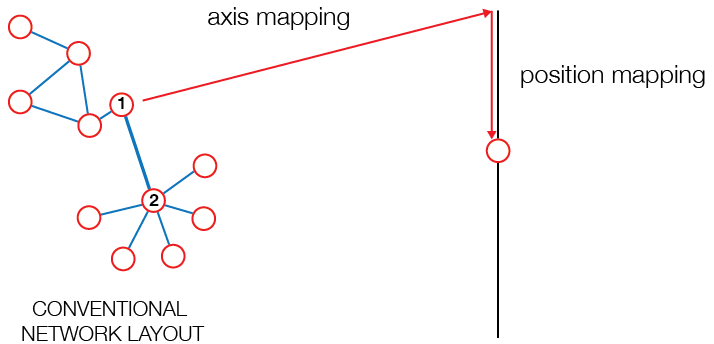
placement of nodes in the linear layout is informed by connectivity and/or annotation

/ layout is controlled solely by meaningful properties

/ interpretation of the visualization is easy, because the layout rules are based on data properties

/ direct comparison between networks is possible

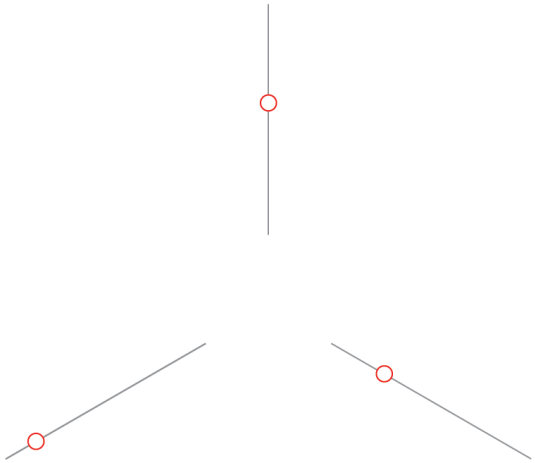
/ describing how the layout was obtained uses meaningful language (*i.e.* based on data properties not aesthetics)



Nodes are mapped and positioned on axes based on structural characteristics and/or annotations. The mappings are meant to be informed by properties of interest, creating a layout that directly illustrates meaningful aspects of the data set.

LAYOUT BASED ON STRUCTURE AND FUNCTION

NODE TO AXIS

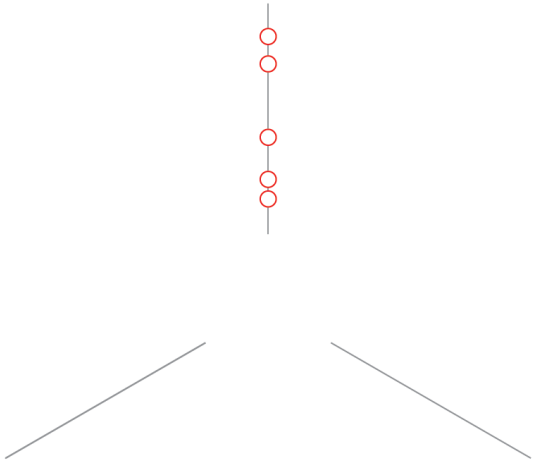


node type (source, sink, both)

node annotation class (*e.g.* gene classification)

AXES CATEGORIZE NODES (NOMINAL SCALE)

AXIS NODE POSITION



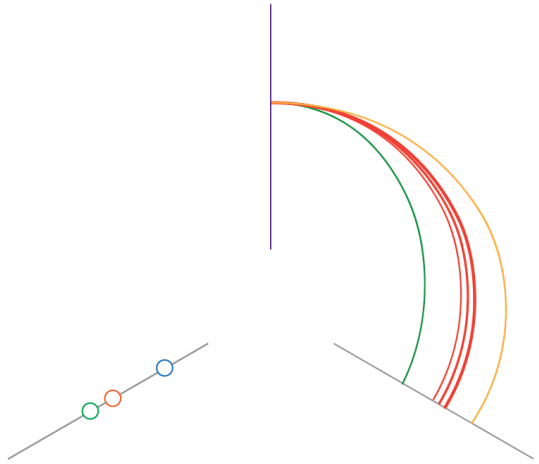
absolute or rank ordered node connectivity

neighbour connectivity

annotation property (*e.g.* expression level)

NODE POSITION ENCODES LOCAL STRUCTURE (ORDINAL OR INTERVAL SCALE)

COLOR AND SHAPE



edge color and transparency controlled by edge weight

glyphs or color codes at node positions classify nodes or layer additional data

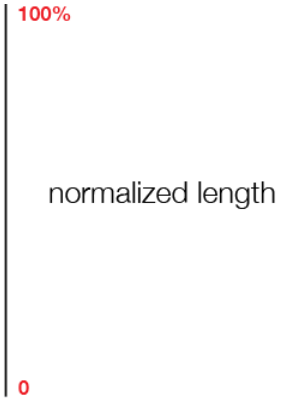
SCALE, ORIENTATION AND SEGMENTATION

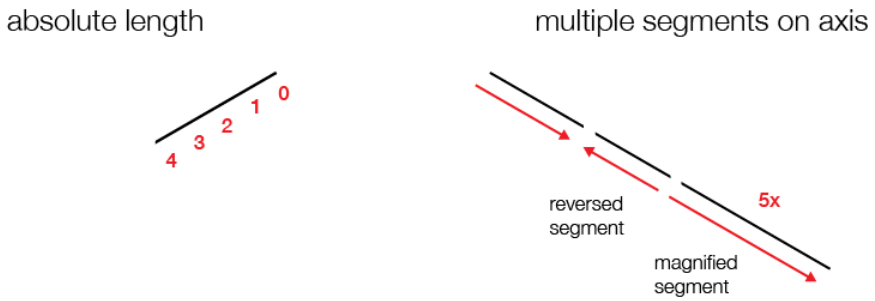
axis subdivision, scale and orientation can be adjusted to add texture and reveal patterns

/ axis length can be absolute (e.g. number of nodes on axis), or normalized

/ an axis can be further divided into segments to further classify nodes (e.g. expression state)

/ individual axes or segments can be reversed or scaled





Each axis may have modfied length, orientation, scale and segmentation.

APPLICATION

Yan et al.[1] compare *E. coli* gene regulatory network to the Linux kernel function call network. The linear layout method presented here greatly facilitates in the visual assessment of differences between these networks.

/ the networks are directional (geneA regulates geneB or functionA calls functionB).

/ nodes are classified based on in/out degree

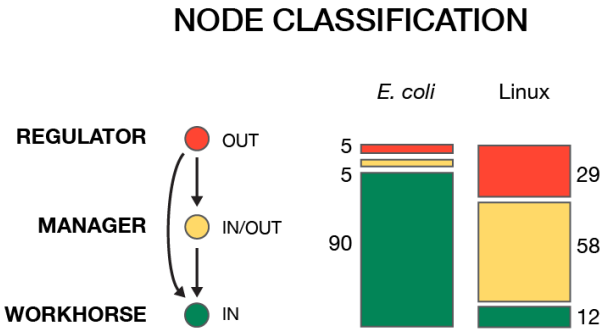
out only (source) – regulator

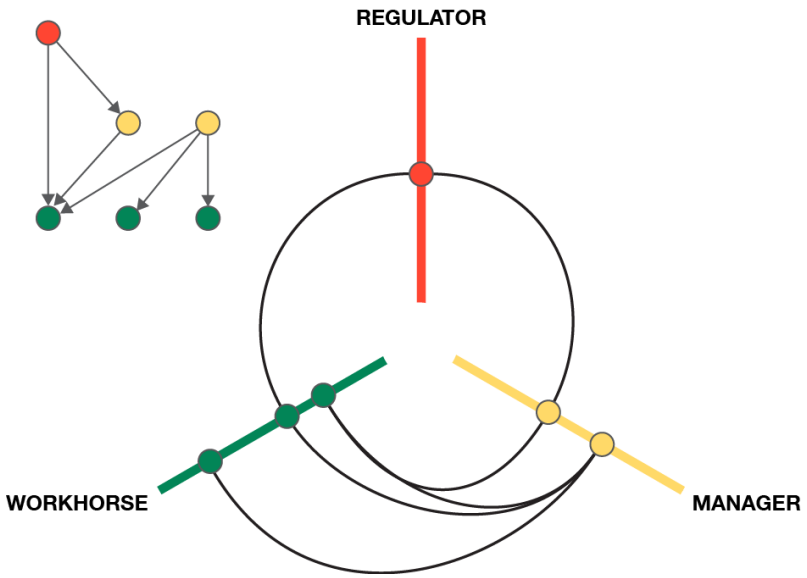
in/out – manager

in only (sink) – workhorse

/ node-to-axis mapping uses this node classification

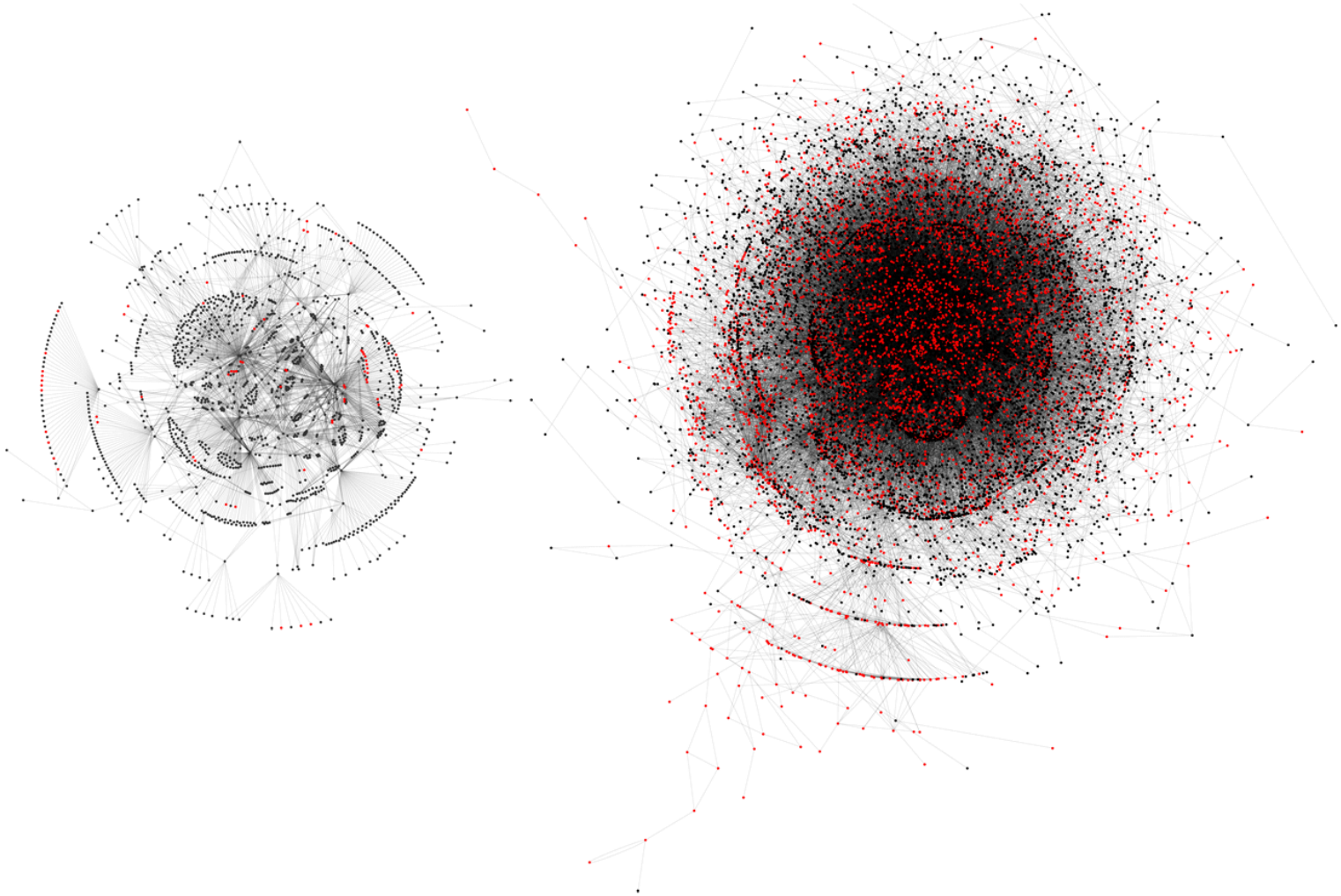
[1] Yan KK, Fang G, Bhardwaj N, Alexander RP, Gerstein M. 2010. Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks. *Proc Natl Acad Sci U S A* ***107(20):*** 9186-9191.





Nodes in the network are classified as regulators (out only), managers (in/out) and workhorses (in only). The *E. coli* network is bottom-heavy (many workhorses), whereas Linux is top heavy (many regulators and managers).

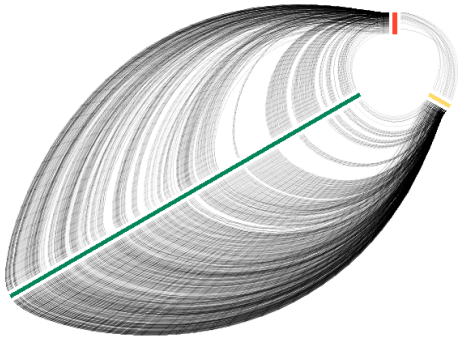
CONVENTIONAL COMPARISON



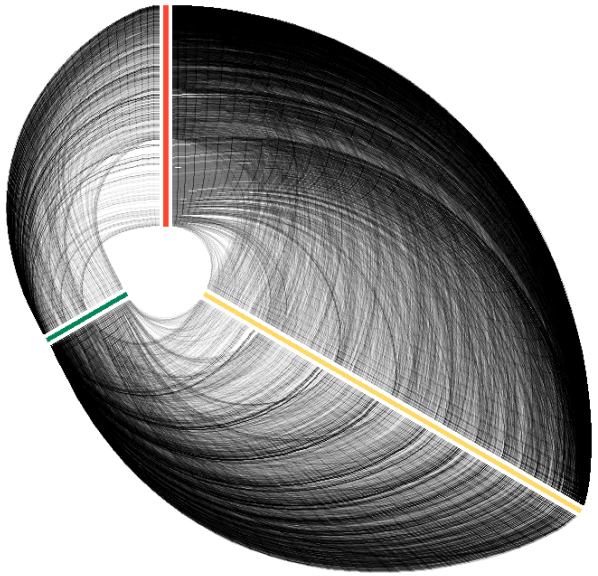
Conventional layouts are not helpful in determining structure of the *E. coli* (left) and Linux (right) networks. Even though the networks are vastly different, except for the network size, all properties are opaque.

LINEAR LAYOUT

Nodes are assigned to axes based on connectivity. Node position is based on rank order of the number of edges at a node (degree). Axis length is proportional to the number of nodes on the axis.



E. coli (6x magnification) / The length of the workhorse (green) axis demonstrates an over-representation in this category. Very few regulator-manager (red-yellow) connections exist. Workhorse connectivity is uniform.

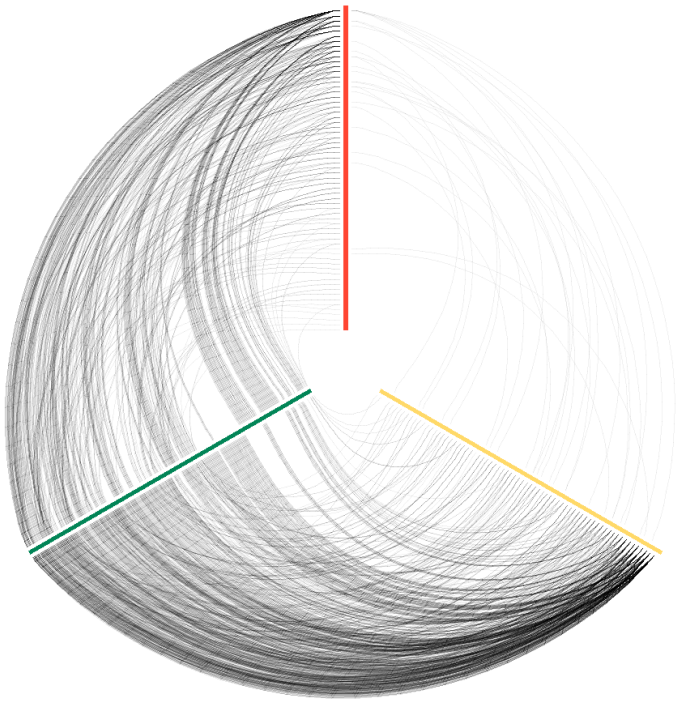


Linux / Large number of regulator-manager connections. Small number of workhorse nodes have very high connectivity (increased density of edges at end of workhorse axis). Approximately 1/3 of the regulators (red) have high connectivity to about 5% of the managers (orange), as evidenced by the converging edge density between the two axes.

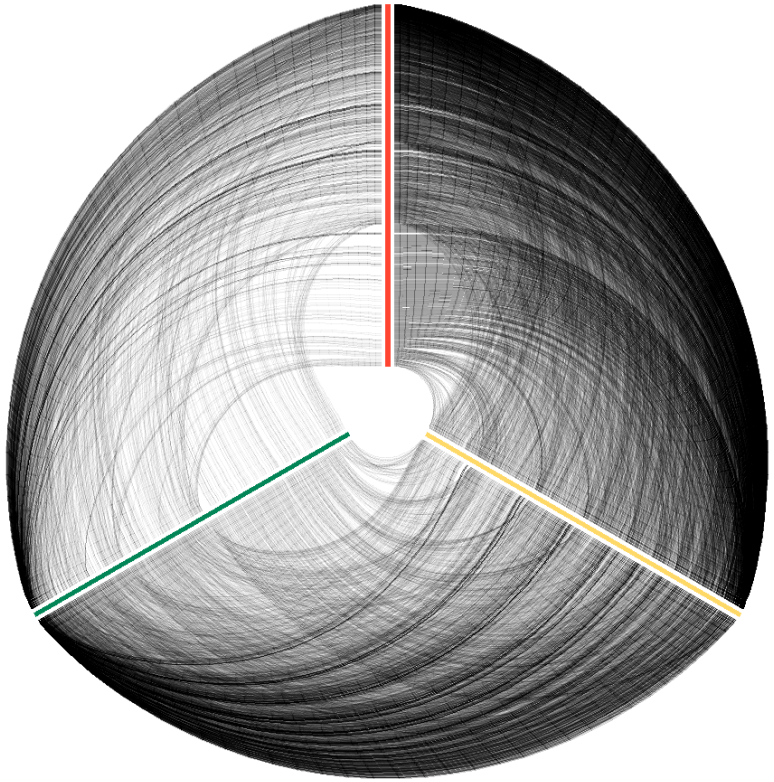
BC Cancer Research Center
Genome Sciences Center

NORMALIZED AXIS LENGTH

The layout method is the same as in the previous slide, but here axis length is normalized to decouple node category size from connectivity patterns. This view allows direct comparison based on node category fractions.



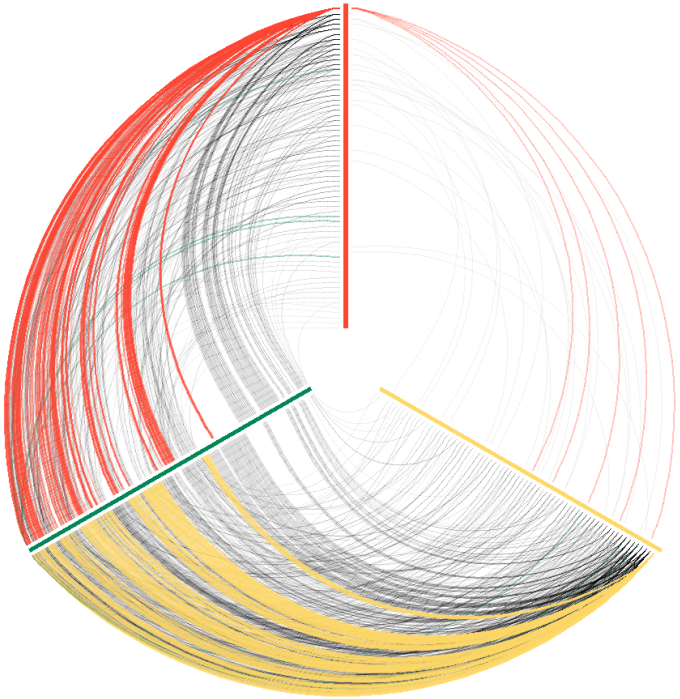
E. coli / Small number of managers are highly connected.



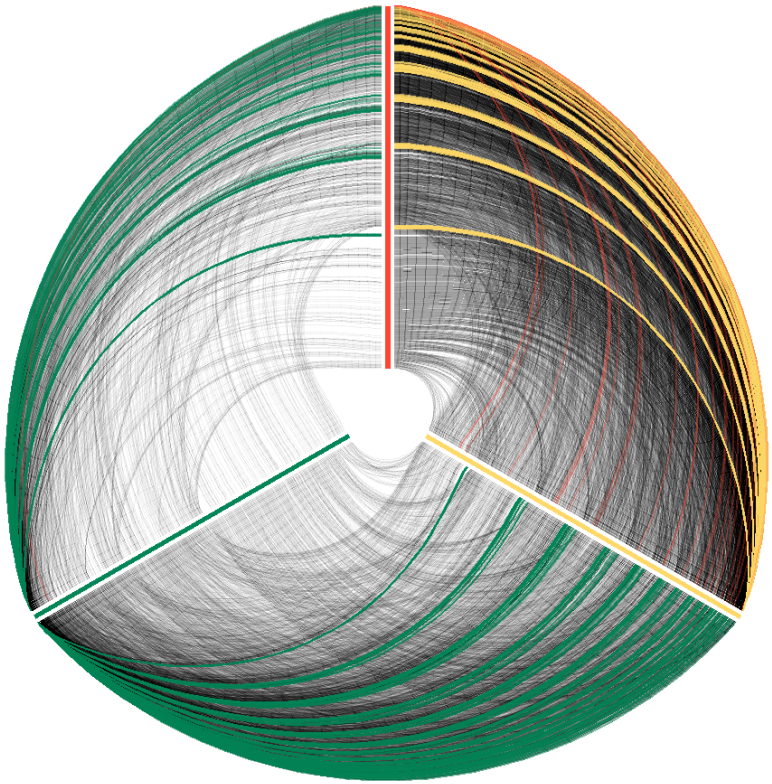
Linux / Heavily connected workhorses are more clearly evidenced when axis length is normalized.

STRUCTURAL ANNOTATION

Layering structural information is easily done using color. Here, links to the most connected node in each group (i.e. most connected regulator, manager, workhorse) are colored by the node's axis color, demonstrating neighbour connectivity around a node category's most connected member.



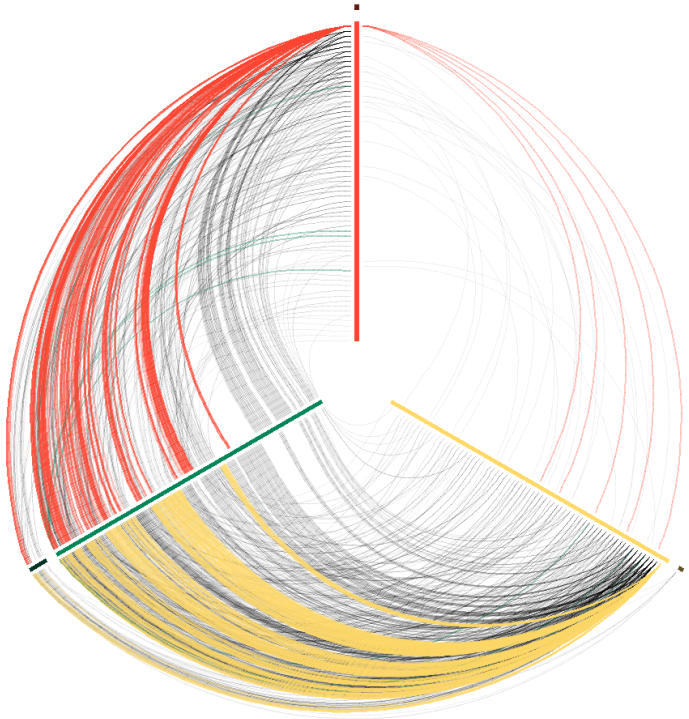
E. coli / The most connected regulator (red) primarily connects to workhorses, but also 5 distinct managers. The most connected manager connects to workhorses.



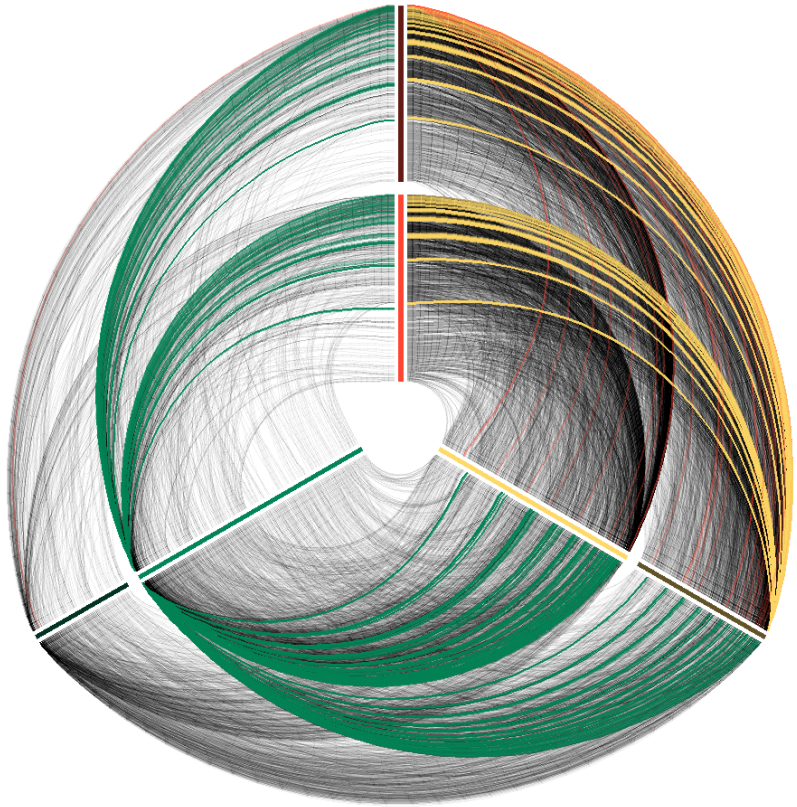
Linux / Unlike *E. coli*, here the most connected manager is connected to regulators. Note the regular banding pattern in the links, suggesting substructure.

SEGMENTATION

Yan et al. further classified each node as either non-persistent or persistent. This is shown here by splitting each axis into two segments that correspond to these two classifications.



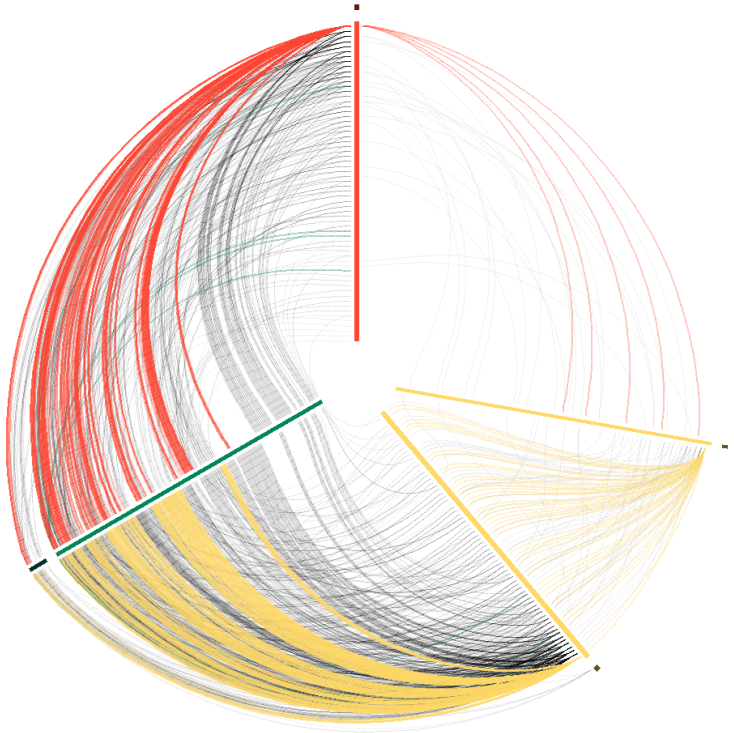
E. coli / Relatively few nodes are classified as persistent (outer segments on each axis).



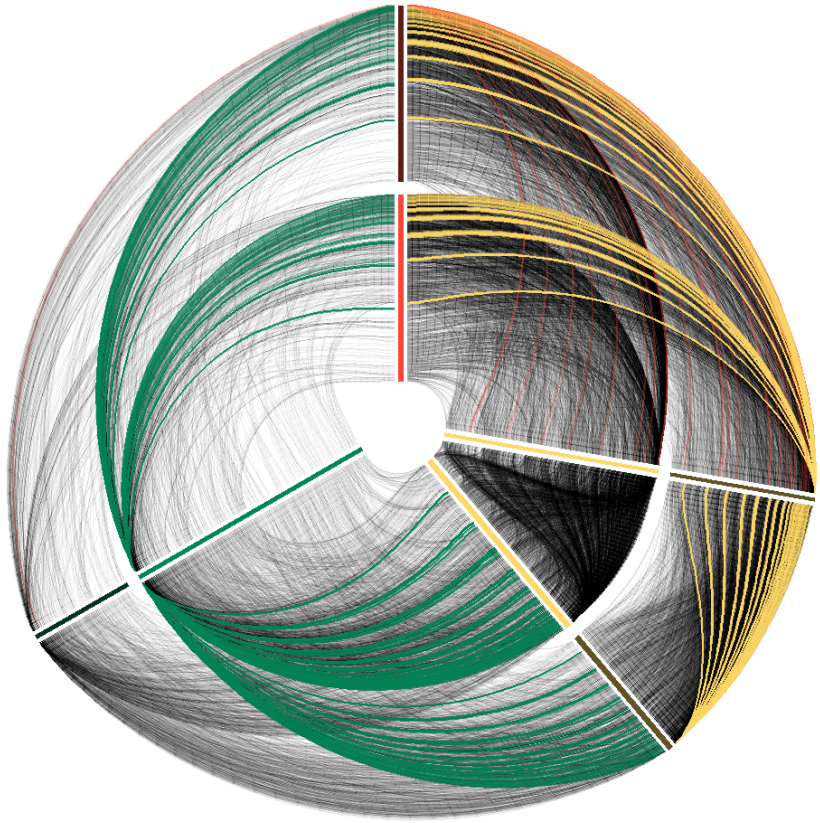
Linux / Each axis contains a near-equal mix of node types. Note that the most connected workhorse is non-persistent (inner segment), whereas the most connected manager is persistent (outer segment).

INTRA-AXIS CONNECTIONS

Managers (in/out nodes) can connect to other managers. These intra-axis links were previously not shown, but can be revealed by cloning the manager axis and displaying manager-manager connections between the cloned axes. The network is directional, with the edge direction clockwise between the two axes.



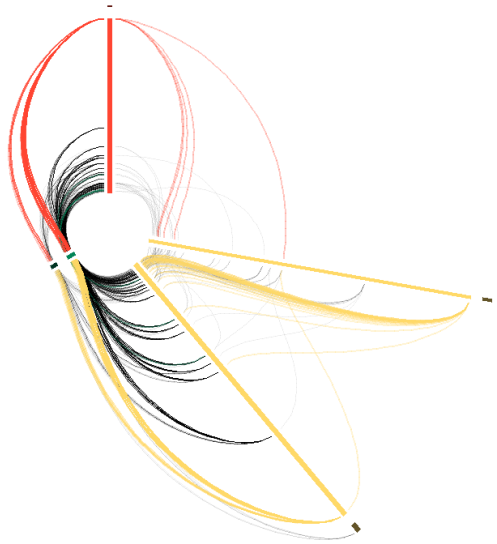
E. coli / The manager-manager connections are largely composed of the most connected manager (its high degree is due to out edges) connecting to other managers. This suggests a cascade.



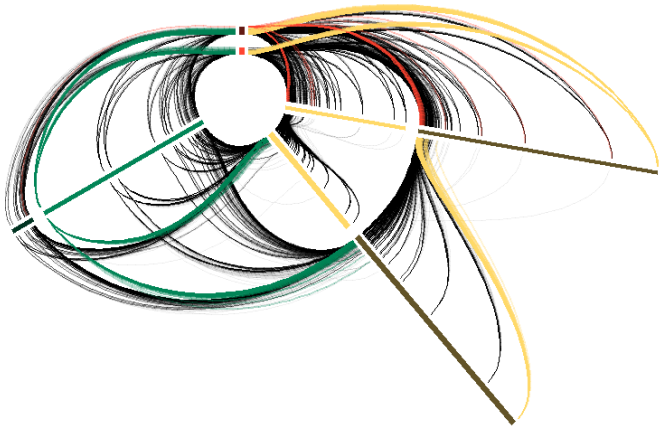
Linux / The most connected manager has a large number of in edges (it's found on the second of the cloned axes, clockwise) and its connectivity to other managers is exclusive to persistent managers.

APPLICATION – ABSOLUTE CONNECTIVITY

Here, node position is based on absolute degree of a node (number of edges). Axis length is therefore proportional to the maximum node degree within a node group.



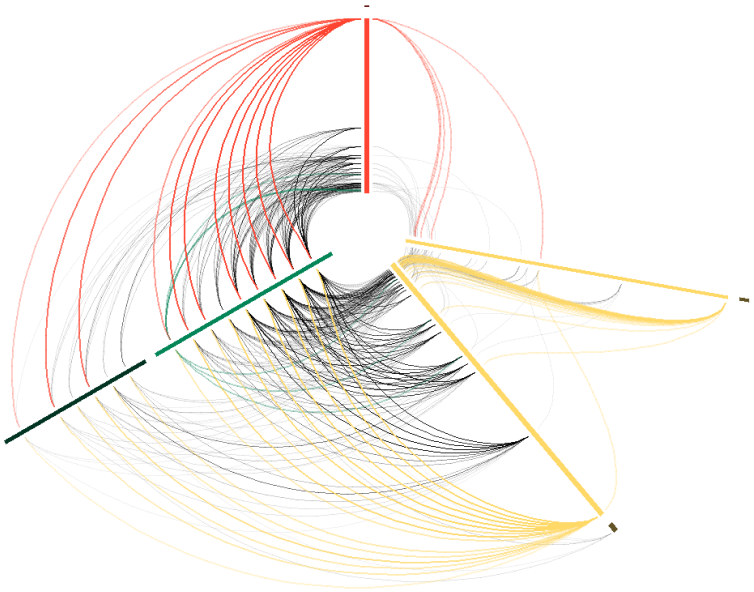
E. coli (3.5x magnification) / The distribution of node degrees becomes evident, with the highest connectivity seen in managers.



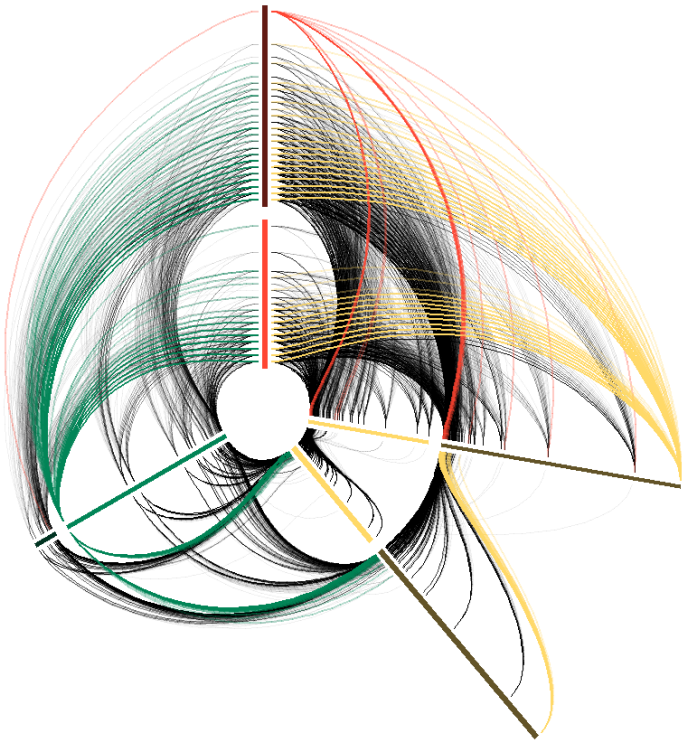
Linux / Intra-manager (yellow) edges reveal that large number of managers with low degree connect to managers with a high degree. From the manager-workhorse links, it is clear that only low degree managers connect to workhorses, whereas high degree managers connect to regulators.

AXIS MAGNIFICATION

Detail can be revealed by magnifying an axis, or individual segments. When the range of node degrees is small, links connect axes at discrete positions.



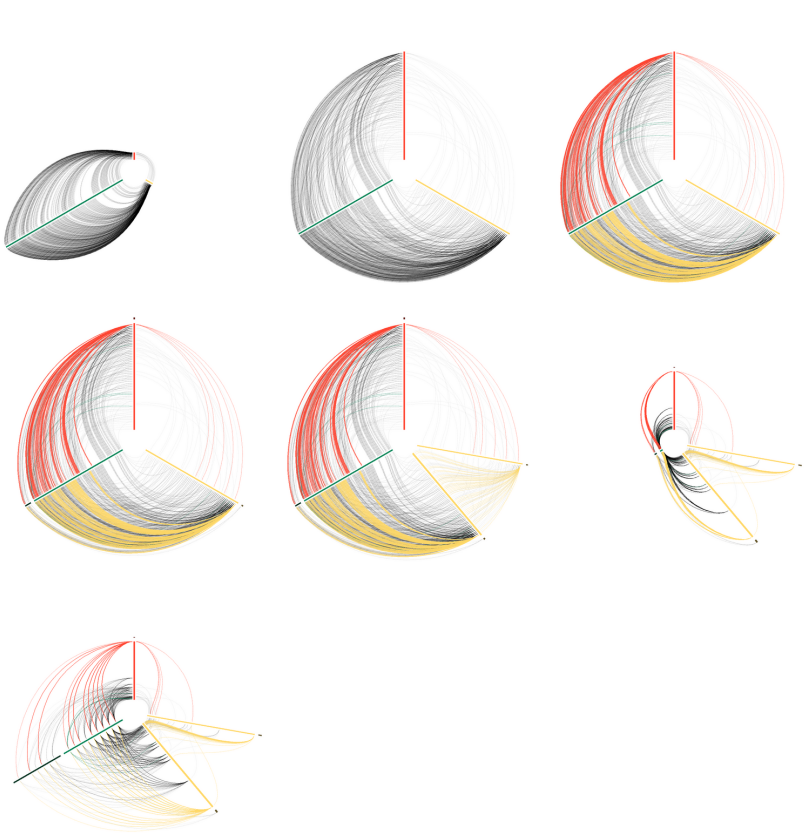
E. coli / Workhorse axis is magnified 25x.



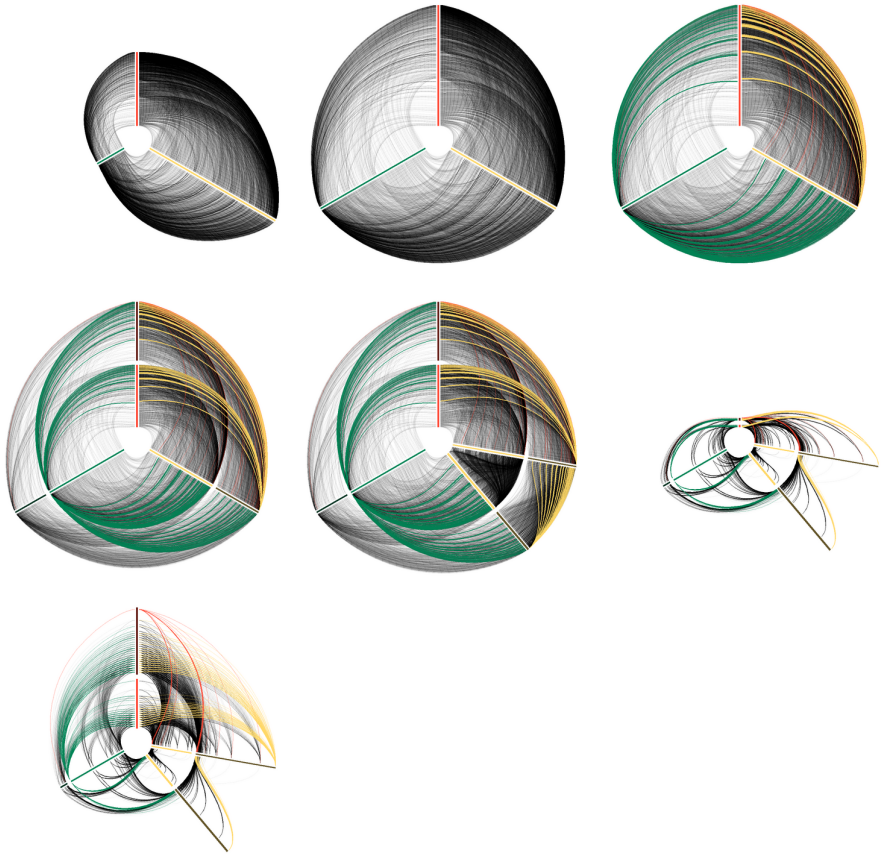
Linux / Regulator axis is magnified 25x.

VISUAL COMPARISON REVEALS DIFFERENCES

Each view reveals different aspects of the two networks, and contrasts distinct differences. Unlike the hairballs, each view is different for the two networks.

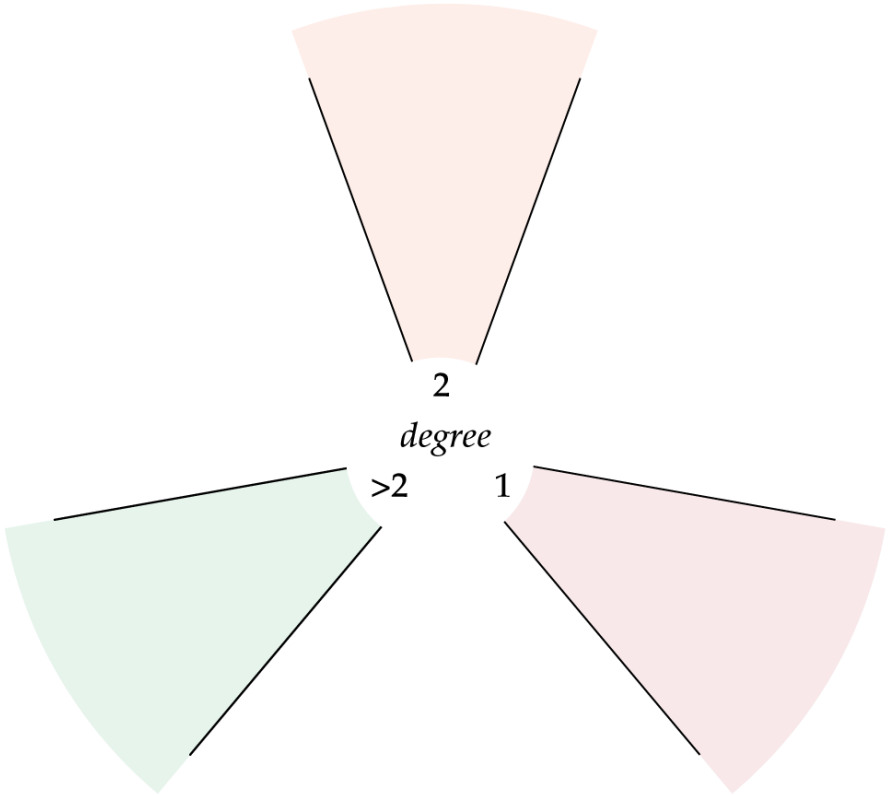


E. coli



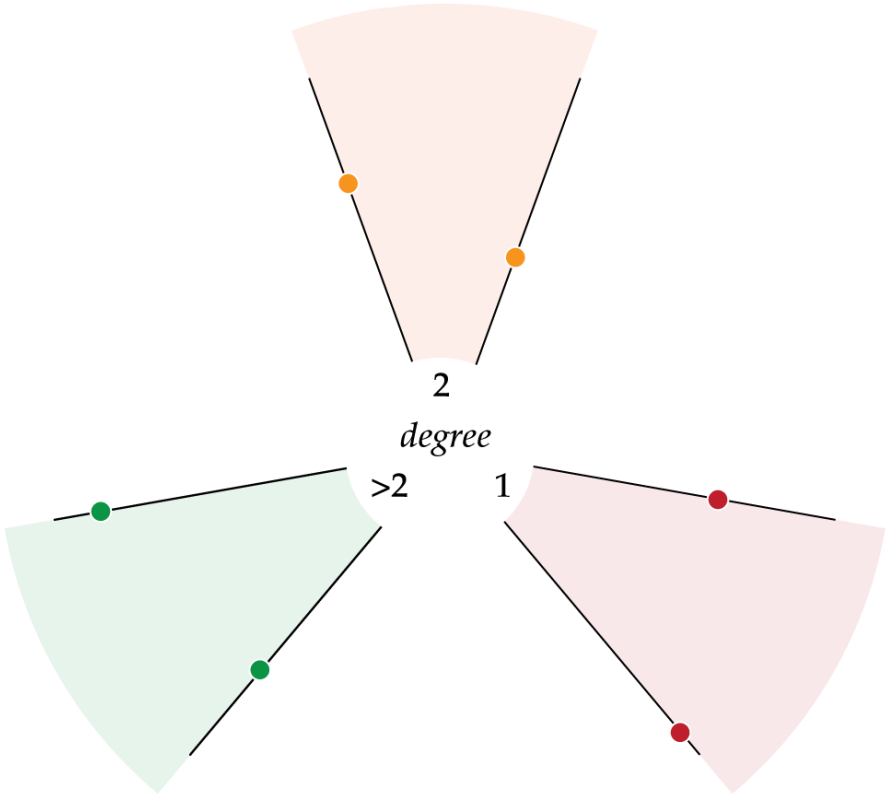
Linux

APPLICATION – UNDIRECTED GRAPHS



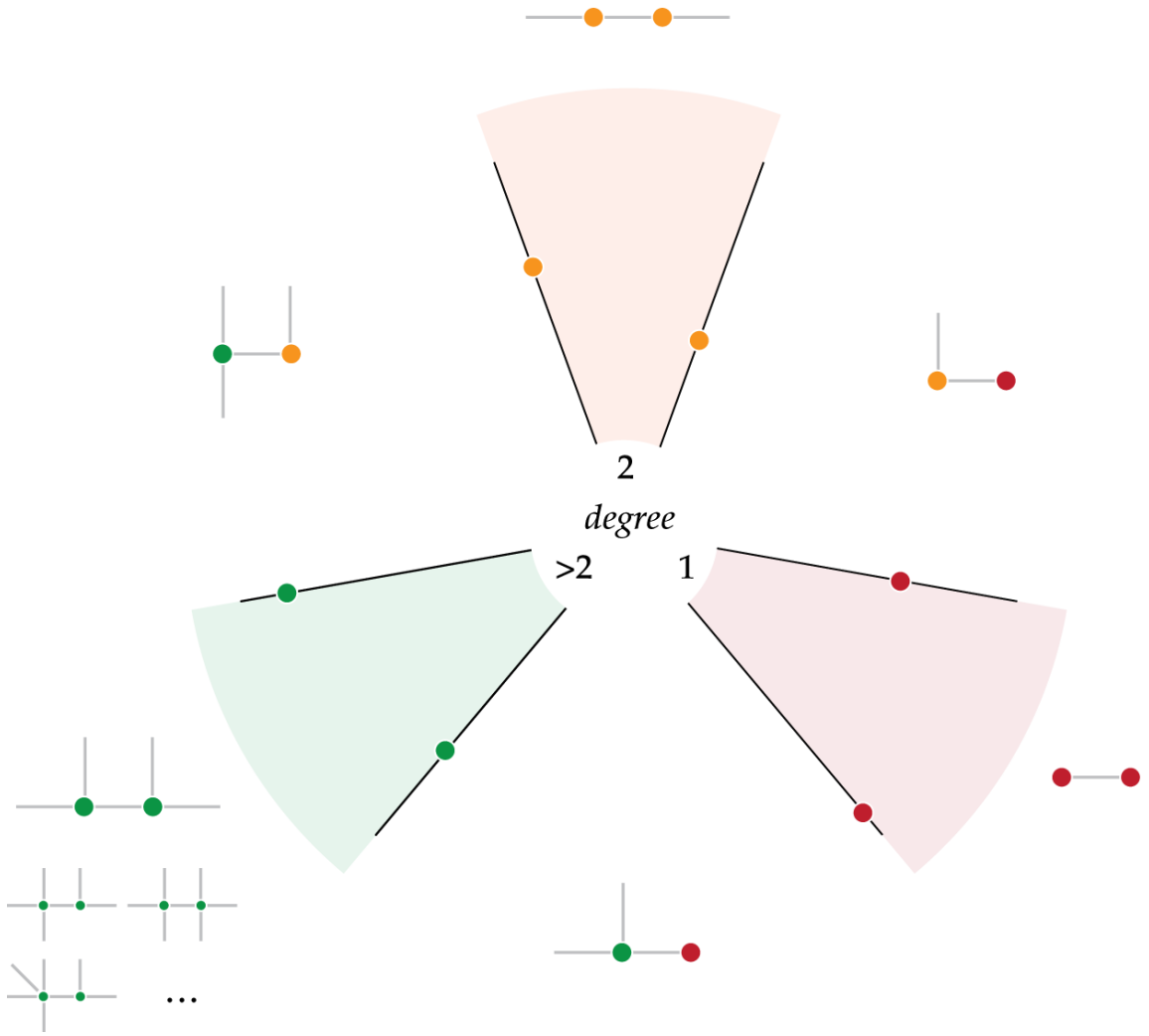
Three axis pairs show nodes with degree 1, 2 and >2.

APPLICATION – UNDIRECTED GRAPHS



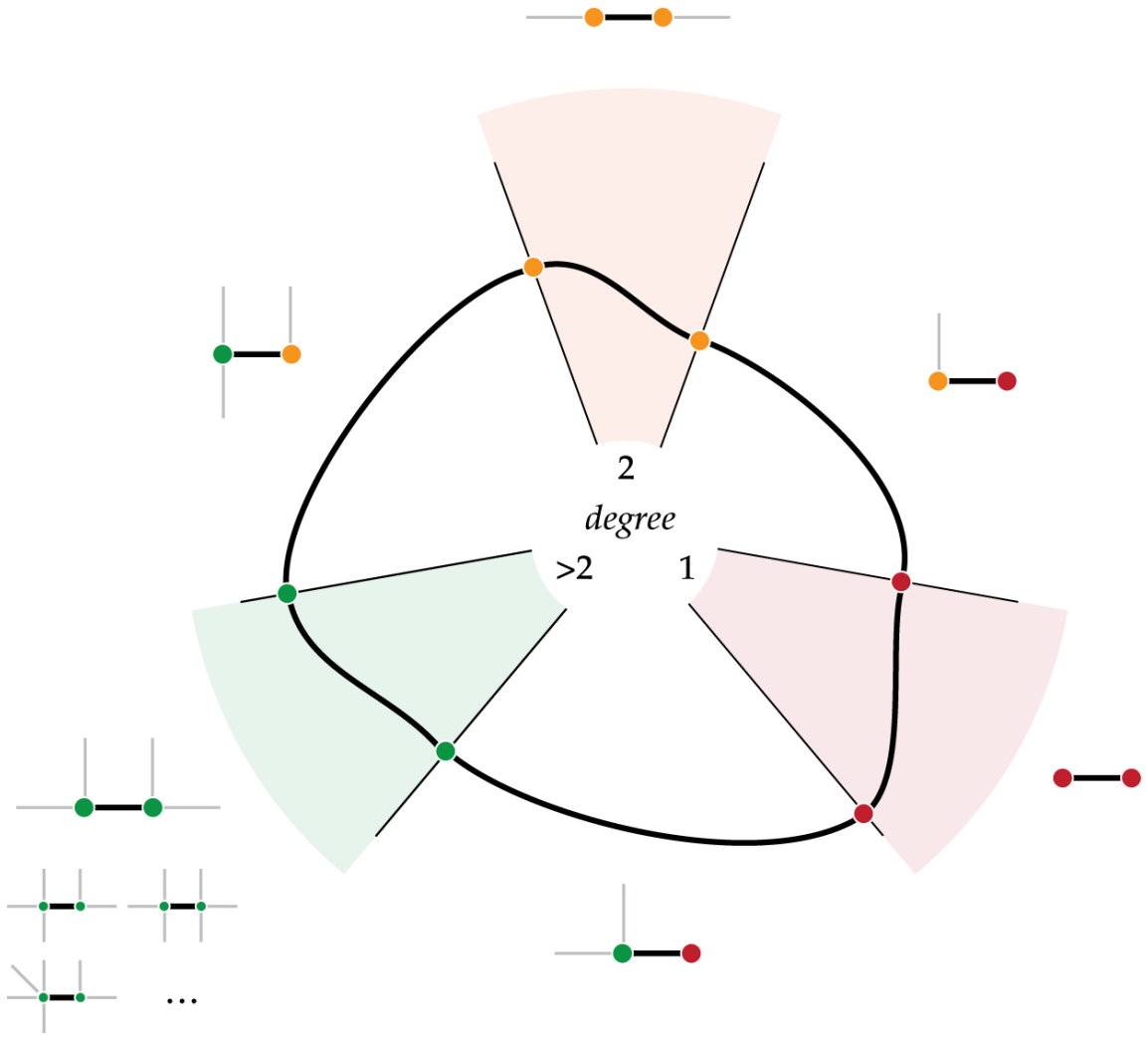
Nodes are placed on the axis based on property of interest, e.g. neighbour connectivity.

APPLICATION – UNDIRECTED GRAPHS



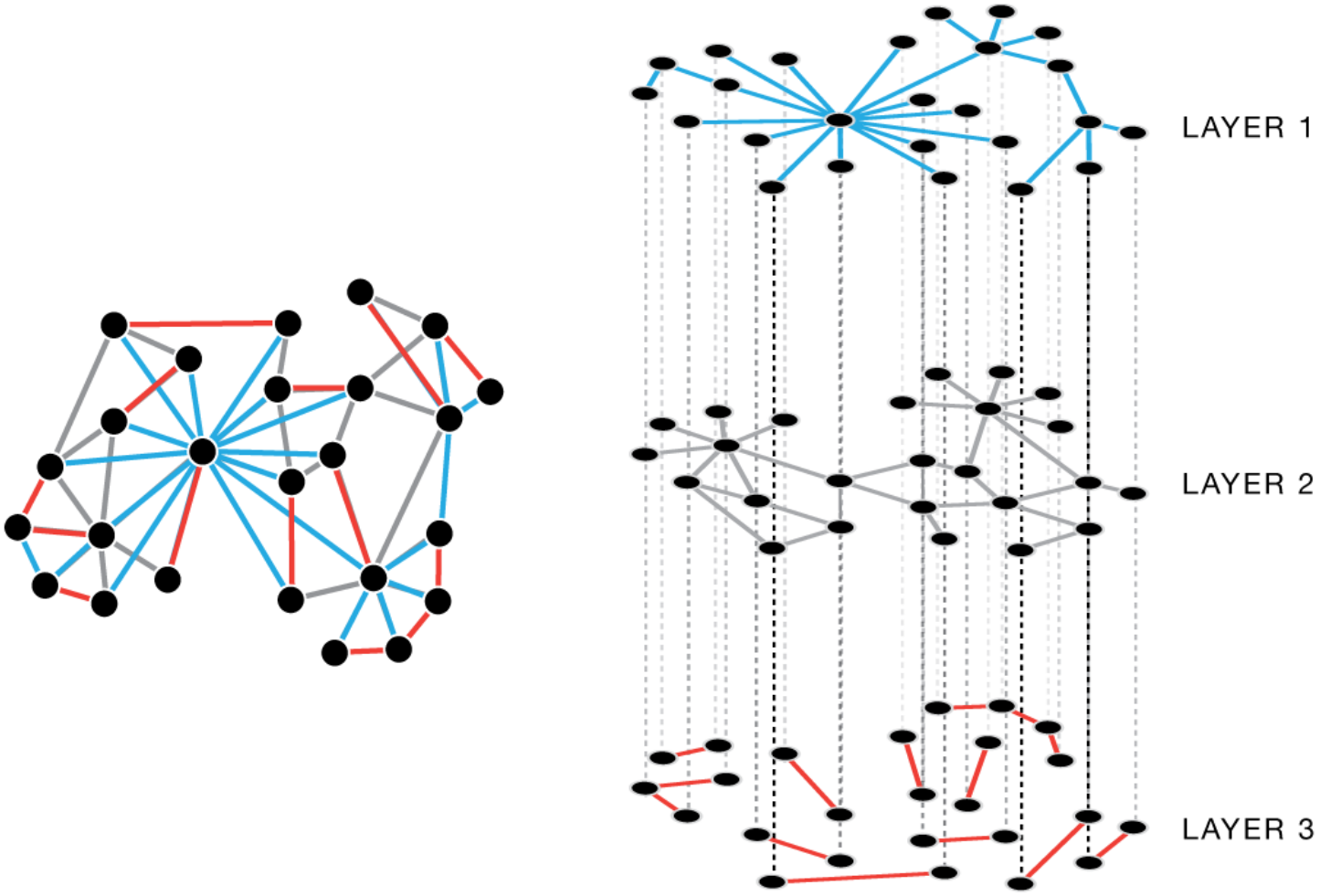
Each of the six regions in the plot are used to represent connections between specific degree pairings: 1-1, 1-2, 2-2, 2-3, etc.

APPLICATION – UNDIRECTED GRAPHS



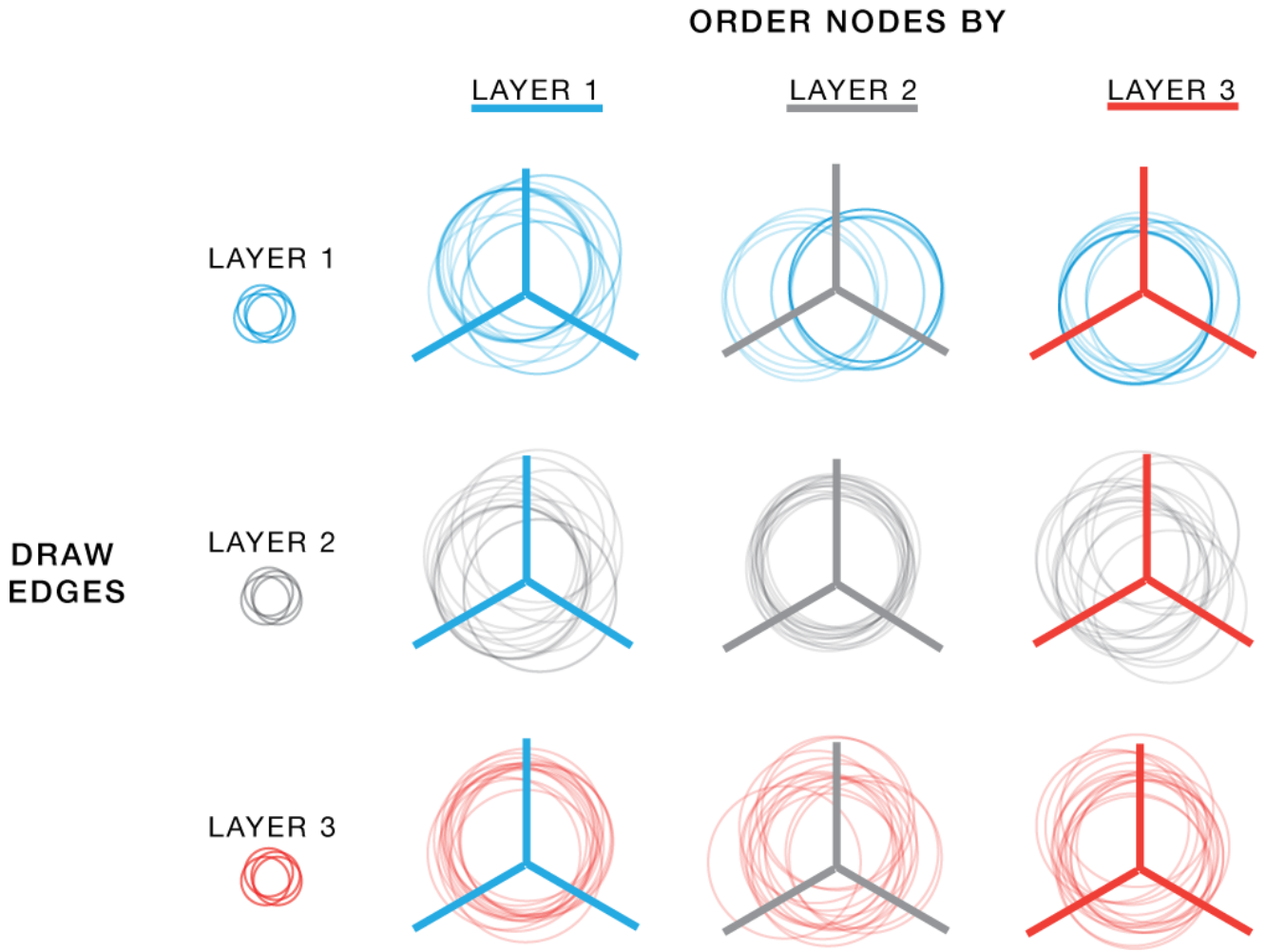
Edges are drawn as curves between the node positions. Edges can be colored by weight, or other meaningful edge (or node pair) property.

APPLICATION – LAYERED NETWORKS



Suppose you have a network composed of three distinct edge groups. These could be thought of as layers of connectivity, with each layer describing a different type of relationship.

APPLICATION – LAYERED NETWORKS



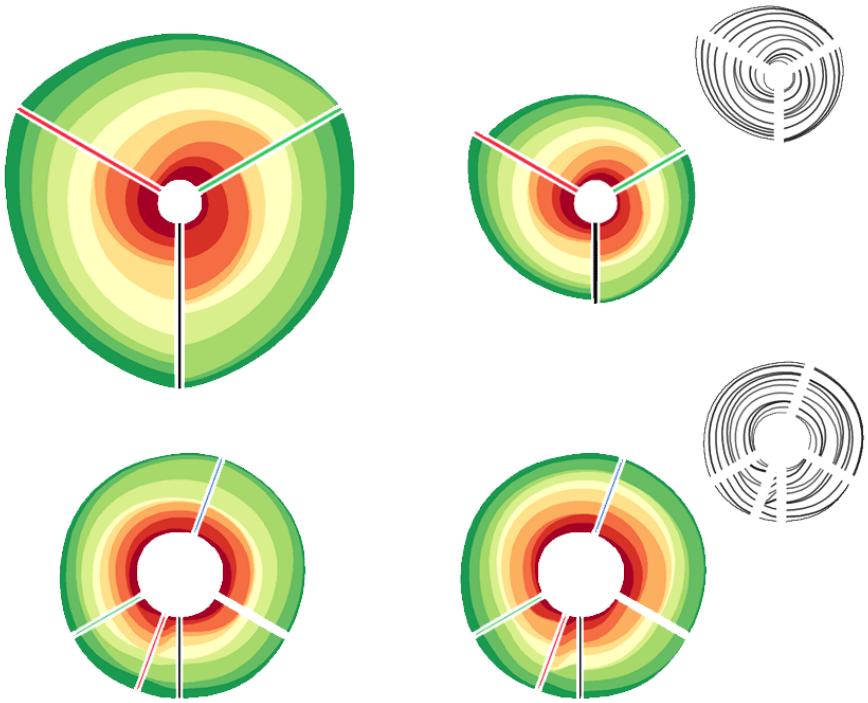
A matrix of linear layouts can reveal how connectivity layers correlate. For each plot the connectivity data that is used to (a) map nodes to axes and determine node position and (b) draw links is *not necessarily the same*.

APPLICATION – STACKED PLOTS

the linear network view can be used to compose stacked bar plots, ideally suitable for comparing multiple ratios

/ edges are drawn as ribbons, with different edge lengths

/ nodes become data values



Application of the network layout to stacked bar plots. The plots are wrapped circularly, creating a comparison loop.

USING THE SOFTWARE

search GIN for “linnet”

/ use local installation

/ download software from web page

mkweb.bcgsc.ca/linnet

ACKNOWLEDGEMENTS

BC CANCER AGENCY

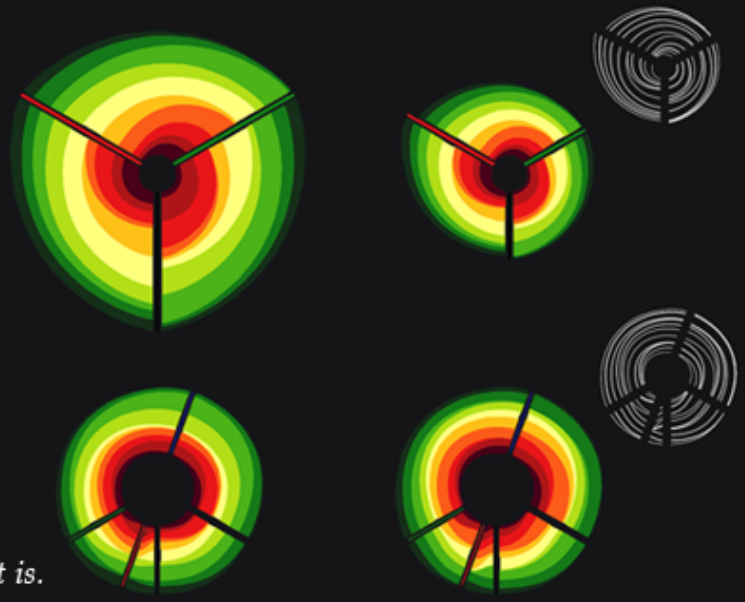
Cydney Nielsen
Shaun Jackman
Rod Docking
Anthony Fejes
Dan Fornika
Jenny Qian

Katayoon Kasaian
Olena Morozova
Inanc Birol
Steven Jones
Marco Marra

MASARYK UNIVERSITY

Martin Lysak

The genius of Gene Rodenberry allowed him to predict a future in which hairballs run amok. In this episode of Star Trek, Trouble with Tribbles, engineer Scott consults with Kirk and Spock about the hairball crisis. Note the tribble in Kirk's cup and those stuck to the walls. It isn't clear how tribbles, which have no legs, can adhere to a vertical surface.
Star Trek Episode 44, 2nd Season, 29 Dec 1967



imagine a world
where something
this pretty
is useful.

we know where it is.

BC Cancer Research Center
Genome Sciences Center