HIVE PLOTS
LINEAR LAYOUT FOR VISUALIZATION OF NETWORKS

DRAWING NETWORKS RATIONALLY

OR

THE END OF HAIRBALLS

MARTIN KRZYWINSKI

Originally presented at Genome Informatics 2010, Hinxton UK (September 17).
WHAT ARE WE HOPING TO ACHIEVE?

Exploring data sets and communicating your findings are two different activities. Typically, the same visualization approach does not suit both.

EXPLORATORY VISUALIZATIONS ARE TOO COMPLEX TO COMMUNICATE.
COMMUNICATIVE VISUALIZATIONS CANNOT BE CREATED UNTIL DATA IS EXPLORED.

The EVEREST PowerWall at Oak Ridge National Laboratory, in Tennessee, is a computer visualization facility. EVEREST stands for Exploratory Visualization Environment for Research in Science and Technology. The 9-meter-wide, 2.4-meter-tall screen can display 35 million pixels of information and is now being used as a tool to model climate change.
http://spectrum.ieee.org/energy/nuclear/slideshow-a-nuclear-family-vacation/0

Comparison of US budget spending by department (left) and related media coverage (right).
http://www.pitchinteractive.com/usbudget/
EXPLORING VS COMMUNICATING

EXPLORE

data driven

hypothesis generating

discover patterns and themes

be thorough by applying variety of approaches

PROCESS AND INTEGRATE

identify minimal set of visual elements to communicate your findings

identify appropriate scale at which differences are shown

COMMUNICATE

theme driven

communicate patterns and themes

be specific in guiding the reader towards your conclusion – the reader is exploring!

EXPLORATORY VISUALIZATIONS ARE USEFUL TO THE EXPERT, WHO KNOWS THE DATA, CAN DISTINGUISH SIGNAL FROM NOISE, AND IDENTIFY INTERESTING PATTERNS.

DESIGNED TO EMPHASIZE CONCLUSIONS, THESE VISUALIZATIONS MUST COMMUNICATE THE STRENGTH AND SIGNIFICANCE OF THE EFFECT. OPTIONALLY, OTHER DATA CAN BE INTEGRATED TO PROVIDE CONTEXT, WHILE PRESERVING CLARITY.
CONSTRaining AMBIGUITY BY IDENTIFYING MEANINGFUL SCALE AND SIGNIFICANCE

RAW  12  54  82  29  25  22  67  61  23  79

MESSAGE ABSENT

DISCRETIZED

COMMUNICATE SCALE

○ 0-30
○ 31-60
○ 61-100

BINNED

COMMUNICATE PROPORTIONS

REFERENCED

COMMUNICATE STRENGTH AND SIGNIFICANCE
EMERGING PATTERNS

Showing the entire data set presents the opportunity to identify emergent patterns – characteristics that are very difficult to express analytically but trivially identified visually.

Nope, no emergent pattern here.

Yup, there’s a pattern here.
Consider how long it took you to identify the form and compare this to how long it would take to create a general program to do the same.
Network data sets are natural inputs for information art. What is information art? A visualization which is beautiful, engaging and compels us to look deeper – it does not need to be informative.
Both are visualizations of a complex system.


Figure 2 and caption quote from Rual et al., Nature 437(7062):1173-8.
A good visualization of the mammoth.

An interaction network is not a tangible physical system like the mammoth and its visualization must be approached by first asking: what questions would I like to answer and what properties would I like to communicate?

A silly question – bear with me. When looking at a scatter plot you know exactly how to identify the meaning of a single data point.
USE LANDMARKS – AXES

The axes provide a reference system.
A COMMON REFERENCE SYSTEM PERMITS COMPARISONS

Scatter plots can be easily compared – such as in this matrix – because they have a common coordinate system. Furthermore, the position of any one point on a plot is directly (and solely) decided by meaningful properties (x,y coordinate).

WHAT IS DIFFERENT? DATA OR COORDINATE SYSTEM?

Our ability to detect linearity is invariant under rotation, but the coordinate system change affects our interpretation of the proportionality in the linear relationship.


Modified figure. Each scatter plot was arbitrarily rotated.
This figure illustrates the challenge posed by an inadequate coordinate system. The data points are too few and too loosely distributed for us to identify meaningful patterns. We’re not very good at identifying loose clusters of patterns against a background of apparent randomness.

LET'S TALK ABOUT NETWORKS

THIS NETWORK

DOESN'T NEED

SOPHISTICATED

VISUALIZATION

THIS ONE DOES
HAIRBALLS ARE IRRATIONAL – THERE IS NO MEANINGFUL COORDINATE SYSTEM

6726 human protein-protein interactions. Rendered with Cytoscape (force directed layout).

4 Nov 2010 / 15
Linear Layout for Visualization of Networks: End of Hairballs / M Krzywinski
SIGNIFICANCE OF SPATIAL PROXIMITY IS ENTANGLED WITH LAYOUT

6726 human protein-protein interactions. Rendered with Cytoscape (force directed layout).
DIFFERENT LAYOUTS – SAME NETWORK?

Subset of the human protein-protein interaction network rendered by Cytoscape using a variety of layouts. 
SAME LAYOUT – DIFFERENT NETWORK?

Subset of the human protein-protein interaction network rendered by Cytoscape. Each visualization uses the same layout (spring embedded), using the previous as a starting point.

SAME HAIRBALL – CAN YOU TELL?

Subset of the human protein-protein interaction network rendered by Cytoscape (edge weighted spring embedded). Each panel shows the same visualization, but with a random rotation and flip.

“The apparent banding pattern of the yellow nodes is an artefact of the graph layout algorithm (Supplementary Data). Importantly, the layout algorithm was not informed by type of supporting evidence and therefore does not explain the evident separation of blue and red edges.”

Figure 2 and caption quote from Rual et al., *Nature* **437**(7062):1173-8
VISUAL CONFUSION

disambiguating layout algorithm from data is impossible

/ a separation of blue and red edges

/ columns of nodes

Authors acknowledge that both effects are artefacts of the layout algorithm.

This suggests the following

/ what are we seeing?

/ what are we supposed to see?

/ is there a better layout algorithm?

Figure 2 and caption quote from Rual et al., Nature 437(7062):1173-8.
WHAT ARE WE HOPING TO SEE? WHAT IS IMPORTANT?

How are these networks different?
### CONVENTIONAL NETWORK VISUALIZATION – PROBLEMS

<table>
<thead>
<tr>
<th>IMPENETRABLE COMPLEXITY</th>
<th>DATA SUBORDINATE TO LAYOUT</th>
<th>COMPARISON IMPOSSIBLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>rapidly grow in visual complexity</td>
<td>hairball’s form is determined by the layout algorithm</td>
<td>layout algorithm is a major influence of the final visualization</td>
</tr>
<tr>
<td>become visually impenetrable</td>
<td>node and edge metadata are subordinate to layout</td>
<td>similar networks may have different layouts</td>
</tr>
</tbody>
</table>

**DEPICTIONS OF LARGE NETWORKS EXCEED RESOLUTION OF OUTPUT AND VISUAL PERCEPTION**

- Important characteristics of the network cease to drive the visualization and cannot be evaluated.
- Hairball visualizations do not clearly reflect aspects of interest.

**DIFFERENCES BETWEEN TWO HAIRBALLS DO NOT NECESSARILY REFLECT A DIFFERENCE IN THE DATA, NOR CLEARLY CAPTURE THE EXTENT OF THE DIFFERENCE**
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.
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
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 modified length, orientation, scale and segmentation.
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

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.
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.
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.
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.
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).
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.
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.
Each view reveals different aspects of the two networks, and contrasts distinct differences. Unlike the hairballs, each view is different for the two networks.
1. confirmed interactions are mostly for proteins with few interactions
2. confirmed self-interaction for most connected mixed protein (TRAF2)
3. more undetected interactions to unconfirmed proteins than unconfirmed interactions to undetected proteins
4. fewer confirmed interactions between unconfirmed than undetected proteins
5. most connected unconfirmed protein has no confirmed or undetected links
6. most connected undetected protein has an unconfirmed link

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.
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*. 
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.
Pairwise comparison between bases in reads, assembly and reference. For example: (A) 20% of reads are unassembled, (B) 30% of reads are unaligned to reference, (C) 2% of reference has no read coverage, (D) 15% of reference has no contig coverage, (E) 60% of reference is constructed by contigs <200kb, (F) there are no contigs >200kb, (G) 20% of contigs are unaligned to reference, (H) 80% of contig bases assembled at k=27.
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 Roddenberry 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.