Visualisation and Analysis of Network Motifs

Weidong Huang, Colin Murray, Xiaobin Shen, Le Song, Ying Xin Wu, Lanbo Zheng
National ICT Australia & University of Sydney, Australia
{weidong.huang,colin.murray,robert.shen,le.song,christine.wu,lanbo.zheng}
@nicta.com.au

Abstract

Many of the complex networks that occur both in nature and in technology are built up from frequently recurring patterns of basic structural elements. These structural patterns known as motifs play a significant role in the function of the network. Visualisation is a useful tool for understanding the structure in a network. The quality of a visualisation can be significantly improved if it effectively displays these motifs. In this paper we present visualisations designed to highlight motifs detected through analysis. We argue that these visualisations designed to show functionally important subgraphs give a greater insight into the function of the network.

1. Introduction

Graph visualisation is an effective means of conveying the structure of relational data. The relational nature of many data-sets makes them ideal candidates for graph visualisation.

Motifs are small, local, patterns of interconnections that occur throughout a network with significantly higher probability than in random networks. These basic structural elements are the building blocks that make up the network and are crucial for understanding how the network functions. Also, a naïve graph drawing approach will not effectively show these motifs. Therefore it is important to design the layout of the graph with these motifs in mind. Typically a network will have a few different types of motifs, each performing a different function. It is beneficial to treat the different categories of sub-graphs differently. Often, these subgraphs are not considered in the layout process and are merely highlighted afterwards. This produces a visualisation that is not as effective at showing these important substructures as it could be. There has been some work that has considered subgraphs in the layout process which we will discuss in section 2.

It is important to perform an analysis on the network and then use this analysis as the basis for visualisation. Analysis of the network allows us to detect the patterns important to the networks function. It is important to perform motif detection prior to designing the visualisation. The visualisation will be more effective if it is designed to convey the properties discovered in the analysis stage.

In this paper, we design a range of visualisations whose primary aim is to emphasise local patterns critical to the networks function while also showing global properties. We present three visualisations, each useful for different situations. The data-set that we use is the transcription regulation network of E. Coli. We believe that by emphasizing the motifs our visualisations will potentially benefit biologists and geneticists. Our techniques could also be useful for any application where small recurring subgraphs are important, for example, social networks.

This paper is organized as follows. In section 2, we review background material. Our network analysis including motif detection is discussed in section 3. In section 4 we describe three visualisation approaches. We then conclude.

2. Background & Related Work

The work of Milo et al. [1] performs motif detection on a range of different networks including food webs, electronic circuits and the world wide web. Also covered is the transcription regulation network of E. Coli. The findings give a greater understanding into how the different networks function. Also, it provides insights into the differences and similarities between different types of networks by comparing motif types across networks. Clearly motifs are very important to the function of a network and should be conveyed in any network visualisation.

Shen et al. [2] also describes motif detection. However, the focus is on a single network. The network investigated is the transcription regulation network of E. Coli. The motifs found by Shen are not exactly the same as those found by Milo. In total four different motifs are found. All are potentially important to the function of the network.

The four important motifs in the transcription regulation network of E. Coli are:

- Feed forward loop (FFL) or triad (Figure 1(a)): a transcription factor regulates a second transcription factor, and then both jointly regulate one or more operons. From an information processing perspective, the main source of information is from the first

transcription factor, while the sink is the operon(s). Genetic information has two choices en route from the source to a sink. It is either relayed in the second transcription factor to ensure a robust control of the operons, or directly passed into the operons to ensure quick information transfer. This motif is also abundant in electronic circuits, with a similar role in that domain.

- Bi-fan Motif (BF) or tetrad (Figure 1(b)): two transcription factors both jointly regulate two operons. The two sinks of the motif have the same sources but due to the difference in the connections from the transcription factors to the operons, the two sinks integrate the incoming information differently. This structure provides two distinct views of the same information.
- Single input module (SIM) (Figure 1(c)): a single transcription factor regulates a set of operons. This structure serves as the distributor of information. The information from a single source is transferred to different operons, possibly changed to influence the next hierarchy.
- Dense overlapping regulons (DOR) (Figure 1(d)): a layer of overlapping interactions between operons and a group of transcription factors. The role of this motif is more complicated. It can be treated as a combination of both BF and SIM as it involves both information distribution and synthesis.

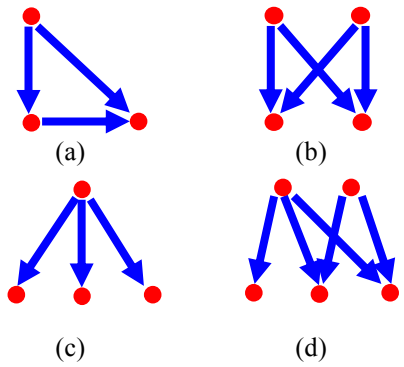

Figure 1 Motifs in the transcriptional regulation network of E. coli. (a) Feed-forward loop (FFL); (b) Bi-fan motif (BF); (c) Single input module (SIM); (d) Dense overlapping regulons (DOR).

The work of Schreiber [3] presents techniques for finding patterns of interest in networks which is the first step in motif analysis. The advantage of this method is that patterns that are not statistically significant or whose function is not known are also considered.

A graph drawing approach has been presented [4] that focuses on the layout of subgraphs. This technique is demonstrated on the transcription regulation network of E. Coli using the SIM motif as the subgraphs. The graph is greatly simplified so that each node represents a SIM and edges represent connections between SIMs. This clearly emphasises the SIMs in the network. However, improvements are possible. The simplification of the graph means a lot of information is lost. No other motif types are visible nor are non-motif nodes. Also, while duplicates of nodes in multiple motifs are highlighted, the connection between a vertex and its clone gets lost.

3. Analysis

In this section we present an analysis of the transcription regulation network of E. Coli. This graph contains 423 nodes and 578 directed edges. The E. Coli network is an information processing network. Nodes represent operons and transcription factors, and directed edges represent interactions. There exist three types of edges. These include activator edges, repressor edges and dual edges. Evidence suggests that the flow of information in such a network is facilitated by its recurring sub-structures [1]. These motifs would not be visible in a naïve approach. We design visualisations specifically to highlight the motifs.

Findings show that feed-forward loop (FFL) and bi-fan motifs (BF) are abundant in this network [1]. These two motifs, though different in their patterns, share their roles in feeding information from one hierarchy to another. We also consider single input module (SIM) motifs and dense overlapping regulons (DOR) as they may also be important [2].

In order to confirm the existence of the above four types of motifs, we have performed statistical tests on the E. Coli network. The algorithm we implemented is similar to that of Milo. Basically, the algorithm scans all rows of the adjacency matrix M of connections between nodes. It searches for non-zero elements (i, j) which represent a connection from node i to node j. The algorithm then recursively traverses the neighbouring vertices connecting vertex i and j until a specific motif is detected.

Due to computational considerations, only patterns with nodes smaller than five are tested. We find that feed-forward loops (FFL) and Bi-fan (BF) motifs are exceptionally rich in the networks, consistent with the findings of Milo et al. Although the existence of SIM and DOR motifs are not as significant as FFL and BF, they may still be important to the function of the network [1,2]. During this analysis we also find a pattern which is very similar to a SIM. The difference is that edges are going in the opposite direction. This motif appears many more times than the SIM in this network. After discussion with the authors of [2], we found out it is a kind of motif which they called the ‘cis-regulatory input function’ [5]. From a visualisation point of view, both of these two kinds of motifs can be regarded as a subgraph with a single node connected to many nodes. In our visualisations we use the term SIM to include both the standard SIM and the ‘cis-regulatory input function’.

4. Visualisation

In this section we describe three visualisation techniques primarily designed to show the motifs of the network. We use the WilmaScope graph drawing tool [6] to produce images of our graph drawings in sections 4.1 and 4.3. The drawings in section 4.2 are produced in Java 3D.

4.1. Clustered drawing

Our first visualisation focuses on showing the two most significant motifs in the network, FFL and BF. This visualisation also has the goal of keeping the graph simple so that it is easy to draw and understand. We would also like to show non-motif parts of the graph as these could still be important.

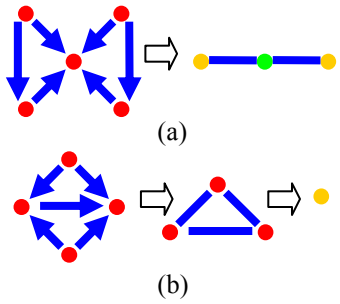

Figure 2 Two collapse processes for (a) hub-grouped triads and (b) twin-grouped triads respectively.

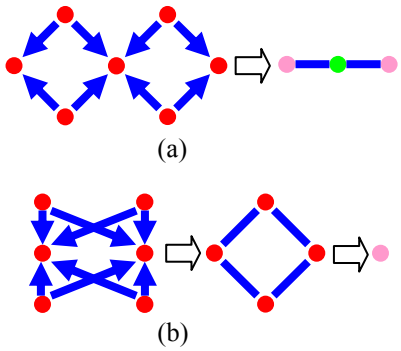

Figure 3 Two collapse processes for (a) hub-grouped tetrads and (b) twin-grouped tetrads respectively.

We achieve these goals by clustering the FFL and BF motifs into supernodes to simplify the graph and make it easier to understand. If a motif does not share a node with another motif, it is simply collapsed into a single node. Otherwise we collapse motifs according to their grouping patterns. To illustrate, we will focus on the case of FFL motifs:

- If two FFLs share a single node (hub-grouped), the common hub is preserved and the remainder of each participating FFL collapse to separate nodes connected to the hub. Non-clustered nodes are coloured red. After this hub-grouped collapse, the common hub is coloured green and FFL nodes are coloured yellow (Figure 2 (a)).
- If FFLs share two nodes (twin-grouped), all these FFLs collapse to a single FFL. If this single FFL does not share a node with any other FFLs, it further collapses to a single node (Figure 2(b)).

The two clustering processes can be generalized to BF groups easily (Figure 3). To distinguish the collapsed FFLs and BFs, collapsed BFs are coloured purple and other conventions remain unchanged.

We now produce a graph consisting of all FFL and BF nodes but no non-motif nodes. We apply our clustering processes to this mixed motif graph and use a force-directed algorithm for the layout.

We then produce a graph consisting of only non-motif edges and their connected nodes. This includes motif nodes which are connected to non-motif nodes causing some duplication. The remaining graph consists of many star-shaped components (SIM motifs).

We want to be able to show the relationship between motifs and non-motifs. Therefore, we combine the collapsed motif drawing with the drawing of the non-motif edges. We use a two-layered layout to accommodate them, the collapsed motifs restricted to the first layer and the non-motif edges restricted to the second layer. This enables us to layout the graph clearly without creating any additional crossings and maintaining some separation between the two subgraphs. Edges are added between duplicates of motif nodes in the non-motif layer and the original node. The resulting graph is showed in Figure 4. From our drawing, it is immediately obvious the role of motif groups in the whole network as well as the connection patterns among motifs. Although some high degree nodes appear outside motifs, many of them have extensive ties with both motif groups and non-motif nodes. High degree nodes are also the structurally most important nodes with almost all motifs groups directly attached to them. These star-shaped hubs are the SIM motifs. The visualisation suggests that they could be very important to the function of the network. Therefore, we will address them in our next visualisation.

The main advantage of this visualisation is that the graph is relatively small. By collapsing motifs we reduce the visual complexity of the graph and this makes it easy to understand. By drawing the motif nodes and non-motif nodes in separate layers it further reduces the visual complexity. It is therefore easier to see the relationship between the two motif types shown (FFL and BF) as well as the patterns between non-motif nodes. Edges that go from one layer to another clearly show the relationship between motif and non-motif nodes.

However, there is a disadvantage in simplifying the graph as the amount of information shown is reduced. It is no longer possible to see the structure of each motif.

Compared to the previous work this improves by showing two types of motif as well as showing non-motif nodes. By leaving nodes that are shared between motifs uncollapsed we also reduce the need for duplicate nodes. These features do increase the complexity of the graph. Our layout is able to overcome this added complexity by separating the nodes into layers. Simple additions such as arrows to indicate direction of edges and labelled nodes further improve its utility. One possible problem is that, even with different colouring, it may be hard to distinguish the two motif types as they are mixed in together.

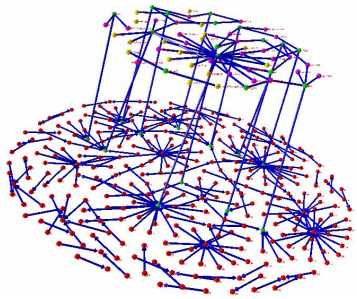

Figure 4 The final layout of our clustered drawing.

4.2. Motif Separation

Our first visualisation had the problem that motifs of different types were mixed together making it difficult to distinguish between them. Therefore, in this visualisation we aim to physically separate the different motif types. While the first visualisation did not focus on the SIM motif, it indicated that it could be quite important. In this visualisation we will focus on the SIM, FFL and DOR motifs.

To separate the different types of motifs we use the idea of drawing the graph on three planes. Each plane represents a motif type. If a node belongs to a motif, then it is restricted to the corresponding plane. Edges between nodes in different motif types are drawn between planes. Nodes occurring in 2 or more motif types are duplicated and an edge is placed between them.

Within each plane nodes are also layered. We use the Sugiyama method modified for 3D graph drawing [7] to minimise edge crossings not only between vertices on the same plane but also between different planes. An overview of the resultant drawing is shown in Figure 5.

Now we discuss the details of additional visualisation features specific to each plane. FFL motifs are drawn on the first plane (red). In an FFL motif, the nodes have different roles, so we use three different colours to tell them apart. Red nodes represent "general transcription factor", green nodes are "specific transcription factors" and yellow nodes are "effector operons". Through the motif detection process, we find the same green or yellow nodes may appear in different FFLs. We do not simply connect all the existing edges, as a lot of crossings may appear, resulting in a poor layout and making it hard to understand. Instead we duplicate those vertices, and use a unique node above or below to connect their clones. We then use the barycentre method [8] to arrange the order of FFLs and the resulting structure is clearer with fewer crossings (see Figure 6). We also draw the motifs congruent to each other to emphasise that they are the same pattern.

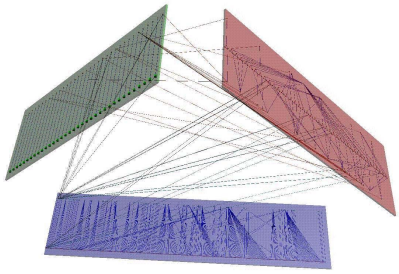

Figure 5 Overview of visualisation with different motif types in different planes.

The second plane (green) is used to draw SIM motifs. To simplify this part of the graph, we collapse the many operons into a single node.

On the remaining plane, we place the DOR motifs as well as any remaining non-motif nodes. Within the plane, we assign the nodes to one of two layers. Sources are in the upper layer and targets are in the lower layer. We again apply the Barycentre method to arrange the order of vertices (see Figure 7).

We arrange the planes to face each other instead of a parallel arrangement. With this approach edges between any two planes will not cross the remaining plane.

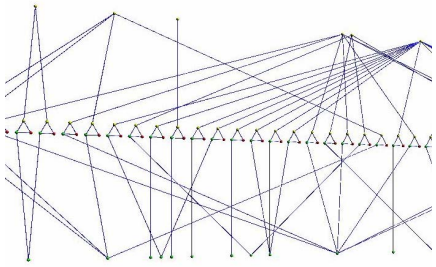

Figure 6 FFLs are small triangles; unique nodes are connecting their clones.

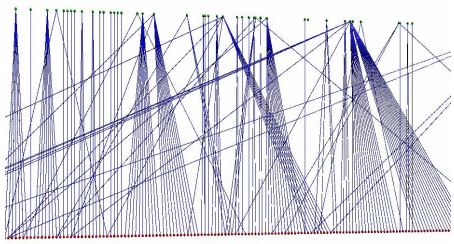

Figure 7 DORs are easy to detect.

This visualisation has a number of advantages over both our first visualisation and the previous work. One advantage is that with the exception of the SIM motifs, it shows the structure of each motif as they are not

clustered. This does have the effect of increasing the visual complexity. However, this is largely overcome by separating the graph into three planes. Within each plane it is fairly easy to see the repetition of small structural patterns that we aimed to identify. Connections between planes may be more difficult to comprehend as those edges are quite long.

Another advantage is that it is easier to distinguish between the different types of motifs as they are physically separated. This visualisation also shows three motif types, more than the previous two visualisations.

While this visualisation does highlight the SIM motifs, their structure is hidden. This, along with the long inter-plane edges makes their contribution to the function of the network difficult to identify.

4.3. Parallel Plane Layout

In this visualisation, we will attempt to overcome the problems of the previous visualisations. We would like to highlight the SIM nodes while still showing their structure. To do this we will not cluster them. In fact, we will show the structure of the entire graph without any simplification at all. We would also like to improve the connections between motifs of different types. To achieve this, we will use parallel planes to allow different motifs types to be close to each other yet still physically separated. To reduce occurrences of edges intersecting planes we will focus on two types of motifs: one is FFL, the other is SIM.

After finding the motifs it was revealed that some separate motifs share vertices. As we want to draw the subgraphs separately from each other, especially the different motif types, we duplicate vertices so that every subgraph remains unchanged and add an edge between duplicate vertices. In this visualisation we show duplicate nodes in such a way that the viewer can easily identify them.

To separate the motif types we assign the nodes to one of three parallel layers (see Figure 8). SIM nodes are placed on the top layer, FFL nodes are placed on the bottom layer and remaining nodes are placed on the middle layer. The force-directed algorithm is applied to produce a nice layout but nodes can not move off their assigned layer.

We place each motif inside a transparent cluster sphere. This highlights the motifs while still showing the entire graph and allowing the viewer to see the structure of each motif (see Figure 9). To distinguish between the two different types of motif, we used colouring on the transparent cluster spheres (see Figure 8). Feed forward loop motifs were highlighted with green cluster spheres while single input module motifs were highlighted with yellow cluster spheres. As the cluster spheres are transparent we can still see the edges between motifs and therefore the relationships between them. The separation of nodes into layers based on motif type further aids the viewer in identifying the motifs.

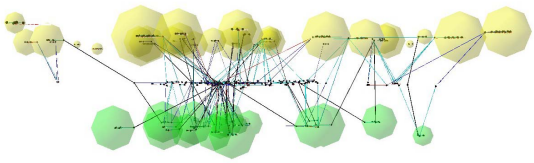

Figure 8 Layering of Nodes

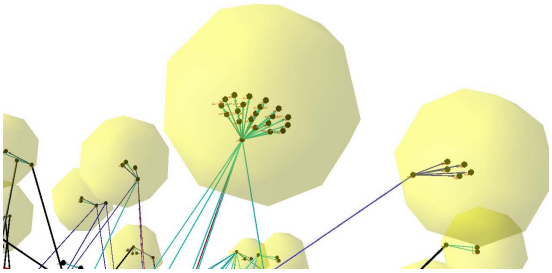

Figure 9 Transparent Spheres highlight the motifs while still showing their structure.

To clearly identify duplicate nodes we place an edge between each node and its duplicates. We visualise this with an undirected, thick black edge (see Figure 10). The edge is made thicker than other edges to stand out and be easily identifiable for the viewer. It is made the same colour as the nodes so that a path consisting of black edges looks like a single entity representing the one operon or transcription factor rather than several nodes.

Edges are represented by arrows. We use colour to distinguish between the different types of edges. Activator edges are coloured light blue, repressor edges are coloured blue and dual edges are coloured red (see Figure 10). This allows the viewer to immediately know what type an edge is based on its colour.

We also display the names of the nodes with labels. This helps to make the graphs useful for analysis.

This visualisation provides a global view by enclosing each motif inside a transparent cluster sphere. From a distance the viewer can look upon the graph as if the cluster spheres are nodes. Our visualisation also provides a local view as the viewer can examine the contents of each cluster sphere. Also, from a distance the structure of the SIMs are indicated by the size of the cluster sphere. We also ensure that it is easy to distinguish between the two motif types by using different colours to highlight the two different types of motif as well as placing different types of motifs on different layers. Our layering approach to drawing the graph is effective in reducing the clutter and produces an easily readable graph without the need to simplify the graph by collapsing motifs. Our visualisation shows all edges between motifs rather than just a simplified graph. This allows for a greater understanding of the relationship between motifs.

This visualisation improves over the second visualisation by clearly showing the structure of SIM nodes as well as their relationship to the rest of the graph. Using colour to identify the different types of

edges also adds to information presented. Edges between different layers are shorter, making it easier to understand the relationships between different motif types. However, there are some edges which travel from the top layer to the bottom, intersecting with the middle layer. Also, we have visualised one fewer motif type in this section.

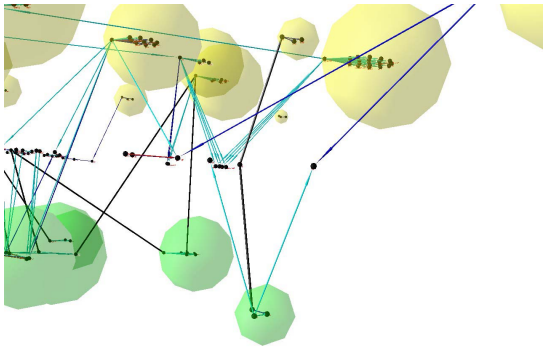

Figure 10 Thick black edges indicate duplicate nodes. The different types of edges are shown by colour.

In the previous work it was difficult to see a connection between a duplicate node and the original node. In this visualisation an edge makes clear this connection without significantly adding to the complexity of the graph.

Overall, this visualisation is quite effective at showing the motifs in the network.

Conclusions

We have presented three visualisations for motif visualisation and shown the results when applied to the transcription regulation network of E. Coli. Our methods improve upon the previous method visualising this network in several ways. The previous method greatly simplified the network, making visualisation easy but conveying little information. Our visualisations provide much more information than the previous method and overcome the added complexity that this introduces by separating the graph into planes. Drawing the graph in two-and-a-half dimensions also provides the best match with the limited 3-D capabilities of the human visual system [9,10]. By providing additional information we provide a greater means for the viewer to understand the presented information.

Our final visualisation shows the most information. However, each of our visualisations have advantages making them useful in different situations. Our first visualisation is useful for showing a simplified overview of the graph. The second is useful for showing many types of motifs and for detailed local views of a single motif. The final visualisation is useful for showing the detailed structure of the entire graph.

Interaction is an important aspect of visualisation that was not covered in this paper but could be investigated in the future. While each visualisation focused on specific motif types, they could also visualize the other motif types, just not at the same time. Interaction could be a useful means for switching between motif types of interest. Also, while the first visualisation provided an overview of the graph, our third visualisation provided more detail. Interaction is also useful for switching between levels of detail.

Another area for future work is evaluation. Without a thorough evaluation, it is difficult to determine which visualisations are most effective.

In summary, a thorough analysis of the data-set is crucial in providing useful visualisations. By performing an analysis into the data, we gained a greater understanding into what the visualisations should focus on. As a result, our visualisations will potentially benefit biologists and geneticists. Our techniques could also be useful for other applications where small recurring subgraphs are important, for example, social networks.

Acknowledgements

National ICT Australia is funded by the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.

References

[1] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002) Network motifs: Simple building blocks of complex networks. Science, 298(5594):824–827.

[2] Shen-Orr, S., Milo, R., Mangan, S. and Alon, U. (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nature Genetices, 31(1):64–68.

[3] Schreiber, F. and Schwöbbermeyer, H. (2004) Towards Motif Detection in Networks Frequency Concepts and Flexible Search. Proc. Intl. Workshop Network Tools and Applications in Biology (NETTAB'04), 91-102.

[4] Gmach, D., Holleis, P. and Zimmermann, T. (2003) Drawing graphs within graphs: A contribution to the graph drawing contest 2003.

[5] Setty, Y., Mayo, A. E., Surette, M. G. and Alon, U. (2003) *Detailed map of a cis-regulatory input function* PNAS, 100:7702-7707.

[6] Dwyer, T. and Eckersley, P. (2001) WilmaScope - An Interactive 3D Graph Visualisation System. Graph Drawing 2001: 442-443.

[7] Hong, S. and Nikolov, N. S. (2005) Layered Drawings of Directed Graphs in Three Dimensions. In proceedings of APVIS 2005: Asia Pacific Symposium on Information Visualisation, 2005. CPRIT 45:69-74.

[8] Sugiyama, K., Tagawa, S. and Toda, M. (1981) Methods for Visual Understandings of Hierarchical System Structures. in IEEE Transactions in Systems, Man, and Cybernetics, vol. smc-11, no.2, pp. 109-125.

[9] Ware, C. (2001) Designing with a 2 1/2d attitude. Information Design Journal 10, 3, 171–182.

[10] Dwyer, T. (2004) Two-and-a-half dimensional visualisation of relational networks. PhD Thesis, The University of Sydney.