Decomposing Social Networks

Whitman Richards
CSAIL- 32-364
Mass. Inst. of Technology
Cambridge, MA 02139
wrichards@mit.edu

Owen Macindoe
Dept. of Computer Science
Mass. Inst. of Technology
Cambridge, MA 02139
owenm@mit.edu

Abstract: Networks having several hundred or more nodes and significant edge probabilities are extremely difficult to visualize. They typically appear as dense clumps, with the various subcomponents completely obscure. We illustrate a method for decomposing a network into aggregates of subgraphs whose topologies are represented as colors in RGB space.

Representing social networks, network scale, visualisation, motifs

I INTRODUCTION

Like many other scientific endeavors, a deep understanding of a phenomena or natural property usually proceeds by a qualitative description of the object under study, followed by attempt to determine the basic parts of the object, and finally the attribution of functions or roles to these parts. Social networks can be studied in a similar manner.

For some time, general observations about networks have been made, such as the probability of node connectivity, the distribution of node degrees, the spectrum, or more specialized measures such as diameter, characteristic path lengths, etc. From some of these measures, interesting properties have been inferred, such as the role of preferential attachment and its relation to scale-free graphs [1,2]. More recently, rather than considering the network as a whole, the microstructure of networks has been explored, identifying very small subgraphs or "motifs" that are common for certain types of networks and those that are rare [8,14,15]. More elusive and difficult has been to identify the larger structural components of networks (i.e. even components with only 25 to 100 nodes.) For example, will such "aggregates" be small networks with their own special topology, or will the aggregates be a cluster of micro "motifs". Moving toward the "aggregate" scale of networks is an important step if functional properties of the network are to be understood.

An analogy may help clarify our approach. In the early analysis of minerals (e.g."rocks"), descriptors such as density, hardness, perhaps shape, etc. were useful. But much greater insight came when rocks were cut and polished to reveal substructures such as dark or shiny aggregates, a variety of crystals, or perhaps threads of veins, etc. With the introduction of spectral tools, including polarization, many components, especially at the finest scale, became colored. The colors were typically robust indicators of the fine structure of the material.

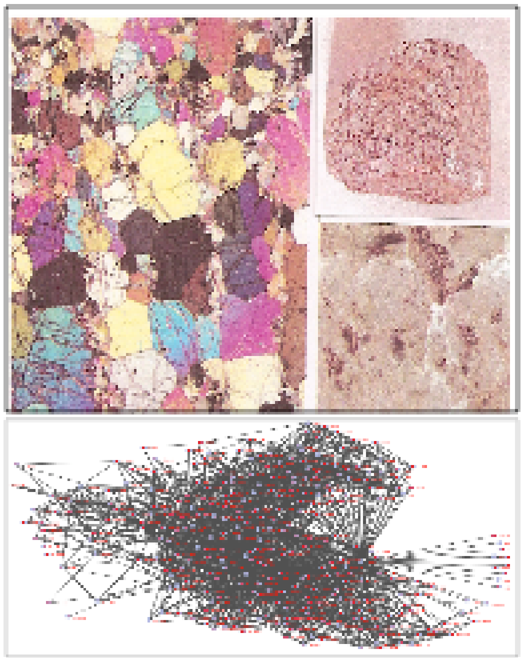

Fig. 1. Bottom panel: a social network of 300 nodes (C. elegans.) [18]
Top right panel: view of a "rock". This view is analogous to viewing a highly connected network, such as the interior of the C.elegans. Middle right: cut and polished section of "the rock" showing hints of aggregates; Upper left panel: the rock seen at a fine scale, illuminated by polarized light to reveal the different crystalline structures.

Our analysis of networks follows a similar path. The intent is not to draw a picture showing "the shape" of a network (as if it were a biological form.) Rather, we attempt to construct a catalog of the ingredients of the network (i.e. rough topologies) and how these topologies change across scales of inspection (see Fig. 1.) Network similarity is thus based on the characterization of the types of subgraphs, not on a visual pattern (which becomes a hopeless exercise for almost all social networks over 500 to 1000 nodes.) Just like spectral lights are used in the study of minerals, so will we provide an index into subgraph topology by colors (in an RGB space.)

II THE BASIC PARAMETERIZATION

Accomplishing our first task – network decomposition — requires a method for parameterizing the topologies of the space of possible subgraphs. Earlier, we introduced a set of three basic, or "atomic" motifs that capture three critical aspects of a strong social effort: leadership ***L***, the bonding of group members ***B,*** and diversity ***D*** [13]. Fig. 2 shows motifs associated with these properties. Leadership ***L*** is captured by the "star" with one dominant vertex of high degree. The bonding ***B*** of group members implies a subgraph topology with high connectivity, with the complete graph K_2 as the limiting case. Diversity ***D*** aims to include small groups of members that are only weakly linked [5]. In start-ups these would be the venture capital members or lawyers, both of whom provide expertise beyond the talents of the founder and the key members of the start-up needed to form the core of product development. Although motivated by concerns for social group structures, evidence has been accumulating that both the star and mesh motifs are significant in social networks, as compared with random graphs [8,14]. Our diversity measure is new, but as mentioned, is related to Granovetter's weak ties, as well as to centrality indicators [10]. If a node has high centrality (i.e. a node joining two dense clusters for example), the network about this node will also have a high diversity. The text box indicates how all three measures are calculated, normalized to the range 0 – 1. Certainly other measures may be invented, and indeed may be found more revealing. However, the ***L B D*** parameterization will illustrate the power and potential of our approach.

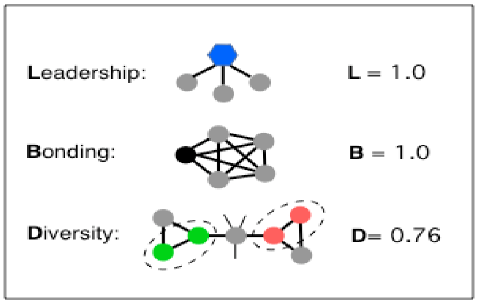

Figure 2. Motifs that capture three key properties of groups, with the maximal or near-maximal values of their associated ***L, B, D*** parameters given on the right. See text box for how the ***L, B, D*** parameters are calculated

III THE SIMPLEX REPRESENTATION

Rather than display ***L, B, D*** values in 3 dimensions, it is convenient to project them onto their <1,1,1> plane as follows:

$$l = L/(L + B + D)$$
$$b = B/(L + B + D)$$
$$d = D/(L + B + D).$$

Figure 3 illustrates this compression. At the top of the Simplex triangle ***B***=1, (green node) which is the position of a dense mesh topology corresponding to the complete graph. All complete graphs, regardless of size will be mapped to this point. Likewise, with ***L*** =1 (***B*** =0, ***D***=0) , the topology of the graph will be a "star", with all star graphs regardless of size

L: For any graph G_n, let d_i be the degree of vertex v_i. The leadership index is then:

$$L = \sum_{i=1}^{n} (d_{\max} - d_i) / ((n-2)(n-1))$$

This relation sums the difference in the degree of a vertex with respect to the maximum degree in G_n, and normalizes this sum by the maximum possible. [4]

B: The bonding index is the number of triangles about that vertex, normalized by the maximum achievable by a graph with the same number of (directed) paths of length 2:

$$B = 6 * (\# triangles)/(\# paths\_length\_two)$$

Note that if G_n is the fully connected graph K_n, then bonding ***B*** is maximal with value "1", whereas for the "star" graph S_n or for any tree T_n, the bonding will be zero .

D: The diversity index counts the number of pairs of disjoint dipoles K_2 in G_n with $n \geq 4$. This count is divided by the number of induced squares in the complete bipartite graph $K_{F[n/2],C[n/2]}$ thus normalizing the measure to [0 – 1]. The square root boosts low ratios: [13]:

$$D = Sqrt[(\# disjo\,\mathrm{int}\_dipoles)/(\frac{1}{2} * \frac{n}{2}(\frac{n}{2} - 1))^2]$$

mapped to the red node at lower right. A ring topology is a simple example of ***D***=1 with ***B*** = ***L*** = 0. More informative, trees with nodes having roughly equal degrees will have high ***D*** values, with low ***L*** ~ 0 and ***B*** = 0.

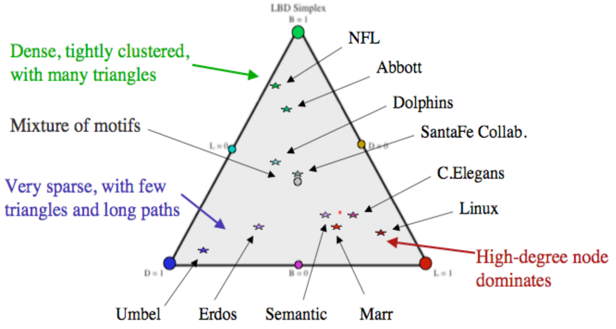

Figure 3. Illustration of the variability of *LBD* indices for different Social networks. For clarity, the raw *LBD* values have been projected onto the 111 plane (ie normalized by their sum.)

Similarly, if diversity becomes maximal, the graph will be located at the lower left. Also shown on Fig 3 are the location of several of the networks we have examined. Note they span most of the Simplex. See Table 1 for parameters of the networks analyzed in this report.

Table 1: Network Parameters

Name	#nodes	edgeProb.	***L B D***	Ref.
Polbooks	105	0.08	.16, .35, .19	11
P3-1 tiling	41	0.1	.22, .29, .20	6
C.elegans	297	0.05	.40, .18, .10	18
Linux08	450	.021	.34, .20, .04	7

IV TWO EQUIVALENT REPRESENTATIONS

From Fig 3, it is obvious that ***LBD*** locations can be encoded as RGB colors, as well as by position in ***LBD*** space, without loss of information. Color has the important property of being an intensive variable. In other words, unlike pictures of graphs that require spatial extent, the color index to a topology requires only a point (or line.) This property means that spatial topologies can be indexed very conveniently as a spectrum of colors. Our mapping is: ***L***-> Red; ***B***-> Green; ***D***-> Blue. When ***LBD*** values are calculated for the network as a single entity, the ***LBD*** color indicates its global characteristic. For example, returning to our "rock" analogy, the color green (***B***=1) indicates the network as a whole must be dense, because edge probability (roughly akin to density) is highly correlated with ***B***. However focusing on ***B*** alone would ignore the crude shape of a network, as indicated in part by the positions on the ***L*** to ***D*** axis. Hence the ***LBD*** color calculated for a whole network is still a good first descriptor of its class.

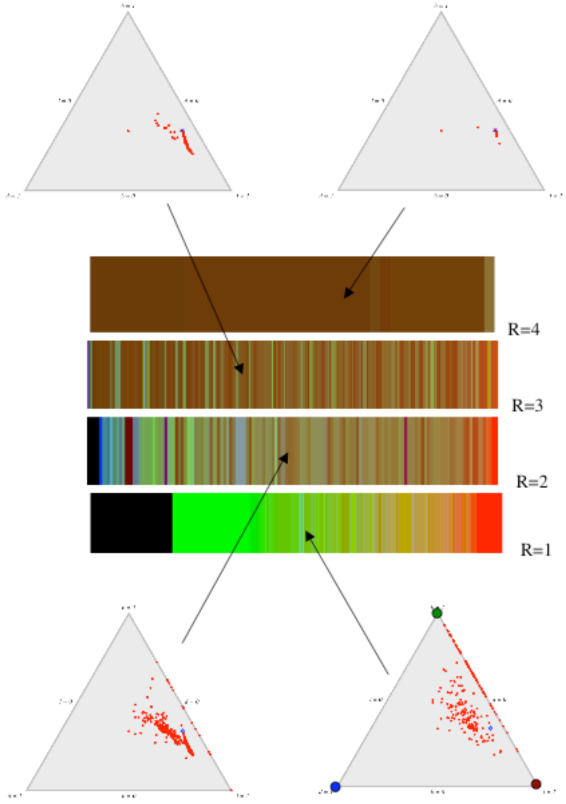

Fig 4.Relation between the ***LBD*** spatial representation and the RGB color form. Each Simplex shows the ***LBD*** fine structure at radius indicated next to the colored panels. (L -> Red; B -> Green; D -> Blue). The data are Linux-08 open source development network.

V NETWORK DECOMPOSUTION

At the extreme opposite to ***LBD*** value for the entire network, we can look at the "crystalline" structure of a network, analogous to the polarized light inspection of the rock shown in Fig. 1. The trick is to recalculate ***LBD*** about each node, using only its neighbors, plus connections between these neighbors. Technically, our subgraphs then have a radius one (one edge step) from the selected node. Obviously we can continue to expand the radius of inclusion of vertices until we reach the diameter of the network. Now all nodes will have the same ***LBD*** value. The lowest color panel and Simplex in Fig 4 shows the "crystalline" structure of the Linux08 network; the top row shows the global characteristics; the other rows are for subgraphs about nodes constructed with intermediate radii.

Our next task is to find aggregates in the network that have similar ***LBD*** values. One might assume that this can be done easily from the spectral bands illustrated in Fig. 4. Specifically, tag each node at some given radius, and find all neighbors with similar ***LBD*** values. If each component of the aggregate indeed had the same ***LBD*** value, this is a straightforward computation – even if one tolerates some minor variations. We use a version of this method in the Appendix to guess an ideal form for a C.elegans aggregate.

Unfortunately, in practice, nodes in aggregates may project to a variety of different nodes outside the aggregate, thus creating a variety of ***LBD*** values, especially for the larger radii about nodes. Furthermore, if an aggregate is heterogeneous, composed of a smaller subgraphs with quite different characteristics (such as a combination of small complete graphs and spidery stars), there will be a "texture" of different colors to be considered – a potentially difficult obstacle (see Macindoe [7] for progress using the ***LBD*** similarity measure.) Classical clustering methods have also attacked this problem, some using K4 or higher subgraphs to quide the clustering [11]

Although one might have a general notion of what should constitute an aggregate of a network, formal approaches need a clear specification. To illustrate, consider the following:

Definition: *an aggregate of a network is a collection of nodes sharing one or more of the following properties:*

(i) ***LBD*** values identical within an epsilon
(ii) Clustering based on latent feature analysis across all scales (i.e. radii)
(iii) A convergence of a "sizeable number" of projections from a well-defined group of nodes, such as a small Kn group of nodes with degree significantly higher than the degree of nodes in the aggregate.

We illustrate the last definition. The intent is to show a novel form of representing network structure, and the advantages of using colors to indicate topology, reserving the spatial form for the manner in which aggregates and subcomponents are related. The procedure is useful especially if the network has a scale-free structure [2].

For scale-free networks, there will be very few nodes of very high degree. Identify say the top epsilon% of these nodes. Determine which are connected. One might find, for example, a complete graph of three such nodes, or perhaps three nodes of very high degree that are not connected. For the PolBooks network shown below (left), we find two unconnected nodes of the highest degree. For C.elegans (right) there is a single isolated maximal node of highest degree, plus two other nodes with runner-up degrees, connected to one another (i.e K2 dipole.) We now proceed to examine the projections of these “top” nodes, and then we will strip these nodes from the network and continue the process until only “scraps” are left.

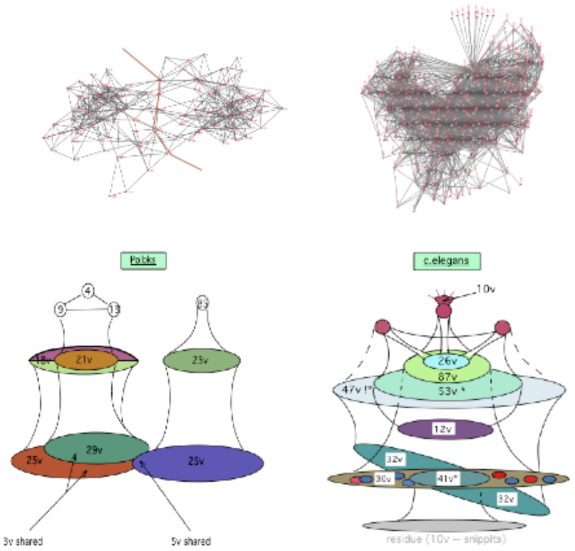

Fig 5. Two networks decomposed by stripping away nodes of lesser degree, using intersections of projections to define aggregate subgraphs. Colors indicate the topological form of the subgraphs. The top nodes were selected by their distinctive high degrees (see text.)

For the PolBooks network at the left of Fig. 5, the vertex degree distribution is (25, 25, 23, 23, 21, 21 …..) In practice, we require a scale-free distribution, with very few nodes of highest degree, well separated from the rest of the pack. (C.elegans qualifies, as will be seen.) However, the small PolBooks network is easily visualized, and illustrates the general method. The four nodes of highest degree are peeled off, with three nodes at the top forming a tight clique K3 (at the left), leaving a single node on the top the right. Consider first the K3 clique. These three nodes all project to one group of 21 vertices, as indicated by the orange colored ellipse. The color of the ellipse indicates the topology of this cluster is a sparse mesh with a few nodes of relatively high degree. At the same level, the two-colored 18v ellipse includes the remaining nodes that are linked to least one but not all of the top K3 members. This cluster has two independent subgraphs: one includes some K4 micro-graphs, the other is sparse with four-cycles.

The top K3 nodes are now stripped away, and the projections from the orange (21v) and purple-green (18v) clusters are explored. These projections are indicated by the ellipses at the next lower tier labeled 29v and 25v. The overlap indicates some shared vertices, namely 3 nodes.

On the lower left of Fig. 5, we now repeat the same process for the remaining component of PolBooks. Here there is a single top node, which is found to project to 23 nodes (green ellipse.) Again, the top node is stripped away, and the projections from the 23v cluster are examined. The deep blue ***LBD*** color for the 25v cluster implies a sparse topology. However, we now see that five nodes in this bottom right 25v cluster are common to the two clusters on the bottom left. This is where the two components of the network come together. The red line in the Polbooks graph shows this cut point.

At the right of Fig 5, a similar decomposition was applied to C.elegans. Here the degree distribution satisfies the scale-free condition: (134, 77, 74, 54, 53….) We peel off the three distinctive nodes of very high degree, shown in red at the top of the decomposition (lower right of Fig. 5.) Two of these nodes are connected, the other (middle) sits alone. As before, we first find all nodes that receive inputs from all three of the top three nodes. There are 26 of these, indicated by the small blue ellipse. The blue color indicates some diversity or sparseness in the connectivity among these nodes. In fact, if we analyze the ***LBD*** values of these nodes, there are two main types with ***LBD*** values as follows: {0.2, 0.3, 0.33} and {0.17, 0, 0.5 }. Note that the second type has no triangles (because ***B***=0) and hence has a “tree” topology, in this case a “chain” of 6 nodes. whereas the first type is crude tessellation. This is an example of a cluster with mixed topologies (and hence technically should appear as a mottled coloring.)

At the same level of decomposition, we have three other ellipses, each corresponding to different patterns of projections from the top three nodes. There are 87 nodes to which the single top red node projects, indicated by the green 87v ellipse. (These do not include nodes in the blue 26v cluster.) Similarly, there are 53 nodes to which both of the two K2 top nodes project, (blue-green ellipse), and the remaining 47 nodes which receive projections form either one of the two top K2 red nodes. Again, as before for the PolBooks network, the top red nodes are stripped away, and the projections from the three (or four) lower tier clusters are examined, The smallest of these is a 12v cluster with only micro-graphs. The more significant aggregates are the green (30v) and olive ellipses (32v), both of which share an additional 41 nodes (dark green ellipse.) At this lowest level of decomposition, the 30v aggregate has many unlinked micro-graphs, indicated by the small colored ellipses.

Returning now to the 41v aggregate which isolates the overlap in the bottom two ellipses, this cluster is marked * because an examination of its topology suggested the corruption of a regular tessellation. In the original studies of C.elegans, the authors [16] noted the presence of "many triangles". Encouraged, we can use our spectral analysis to estimate a possible "ideal" form of the actual neural network, prior to any corruption. This analysis is given in Appendix 1.

Finally, we decompose the Linux08 open source development network. (See Fig 4 for the fine-structure plots.) The breakdown is similar to that in C.elegans, but note that the coloration, and hence the topologies of the aggregate subgraphs are quite different. Especially striking is that at the lowest level, the aggregates are not fully connected, but are broken up into "crystalline-like" micro graphs. These include dipoles, micro-meshes, etc. This breakdown is anticipated by the large number of free ends that appear in the global picture of the network (left in Fig. 6.) To date, we have yet to complete our analysis of this network.

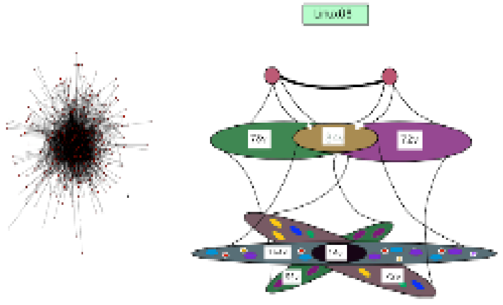

Fig.6 A decomposition of the Linux08 network.

VI FUNCTIONALITY

The third, and most important step in network analysis is assigning function to structure. Not all networks can be expected to exhibit functional properties. For example, the Polbooks network is based on choices of political books by the same individual. The two main clusters correspond to taste differences, such as for conservative vs. liberal viewpoints. With the exception of the "cut point nodes" which join the two clusters, one does not expect to be able to ascribe functional properties to the structural decomposition shown in Fig. 5.

C.elegans, on the other hand, is a creature that is engaged in information processing, taking in sensory inputs, evaluation of these inputs, with decisions as to actions to take next. If there appears evidence for a network tessellation (as in Appendix 1), then there is the hope of assigning function to structure. The development of Linux08 is analogous in this respect [7]. Our decomposition shows clear roles for a small group of "leaders", their immediate followers, plus a host of others who are contributing to aspects of the code development. Although these functional assignments are intuitive, there is the strong belief, as in C.elegans, that different components of the Linux08 network were performing different functional roles. Of special interest to network understanding is whether aggregates with similar topologies in different networks are performing similar functions. If so, we have the beginnings of a network science.

ACKNOWLEDGMENT

Omprash Gnawali provided the Linux data. Comments from Viren Jain and the reviewers were much appreciated. Supported by AFOSR under a MURI grant #A9550-05-01-0321.

REFERENCES

[1] Barabasi, A-L (2002) *Linked: The New Science of Networks.* Perseus Press, NY.

[2] Barabasi,A. & R.Albert (1999) Emergence of scaling in random networks, Science 286, 509-511.

[3] Bollabas, B. (2001) *Random Graphs*, 2nnd Ed. Cambridge Univ. Press.

[4] Freeman, L. C. (1978) Centrality in social networks: conceptual clarification. *Social Networks*, 1: 215 – 239.

[5] Granovetter, M. (1973) The strength of weak ties. Amer. Jrnl. Sociology 78, 1360 – 1380

[6] Grunbaum, B. & G. Shephard (1987) Tilings and Patterns. W. H. Freeman & Co.

[7] Macindoe, O. (2010) Investigations of the fine structure of social networks. MS Thesis Jun 2010 Dept. Elec. Eng & Comp Sci, MIT. (http://people.csail.mit.edu/owenm/netdata.html).

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

[9] Newman, M. E. J. (2003) The structure and function of complex networks. *SIAM Review* 45: 167 – 256.

[10] Newman, M. E. J. (2005) A measure of betweenness centrality based on random walks. Soocial Networks 27, 39 – 54..

[11] Palla, G. Derenyi, I. Farkes, I. & T. Viesek (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814 –818.

[12] Read, R.C. and R.J. Wilson (1998) *An Atlas of Graphs*. Oxford Press.

[13] Richards, W. and N. Wormald (2009) Representing small group evolution. Proceedings IEEE Conference on Social Computing (SocialCom2009. SIN09-232. [Also MIT-CSAIL-TR-2009-012]

[14] Stoica, A and C. Prieur (2009)Structure of neighborhoods in a large social network. Proceedings IEEE Conference on Social Computing (SocialCom2009) SocialCom09-289.

[15] Wernicke, S. (2006) Efficient detection of network motifs. IEEE/ACM Trans. Compt. Biol. BioInformatics 3, 347-359..

[16] White, J. , E. Southgate, J Thomson, S. Brenner (1986). The structrure of the nervous system of the nemaotade C.elegans. Phil Trans. Roy. Soc. Lond. B, 314, 1 – 34.

VII APPENDIX: spectral analysis of C.Elegans

Obviously inferring an ideal topology from any 41 node network is essentially intractable [3, 12] However, we can determine by trial an error if a regular tessellation can be corrupted by random processes to create ***LBD*** values at different scales that resemble those found in the network under study. One network that is appealing is the P3-1 isohedral polygonal tiling (Grunbaum & Shephard, [6] pg 473.) shown in Fig. 7.

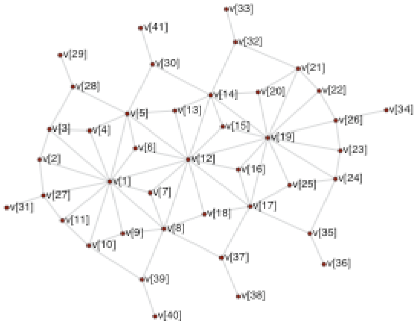

Fig.7 A P3-1 polygonal tiling, slightly modified at boundaries.

This portion of the tiling has 41 nodes with 78 edges. To mimic the C.elegans aggregate of interest, an additional 8 edges were added, linking random pairs, thus making the edge probabilities and node counts the same. Next, 8 edges were chosen at random (uniform distribution over all edges), and one vertex for each edge was re-wired to another in the network.

In the Fig. 8, we exhibit the color spectrum at radius 1 and 2 for three versions of the tessellated network. The middle two panels show the uncorrupted tessellation. Note the broad bands of homogeneous color, which is expected for any very regular graph. The top two panels show the color spectra of this ideal tessellation, with the addition of the random edges mentioned above. These changes would correspond to a 20% corruption for the idea network. The bottom two panels in Fig. 8 show the C.elegans color spectra for the aggregate. Considering that random processes do not give a unique corruption, the similarity between the top two and bottom two panels is encouraging. To first order, the C.elegans network might well be a corruption of the P3-1 tiling.

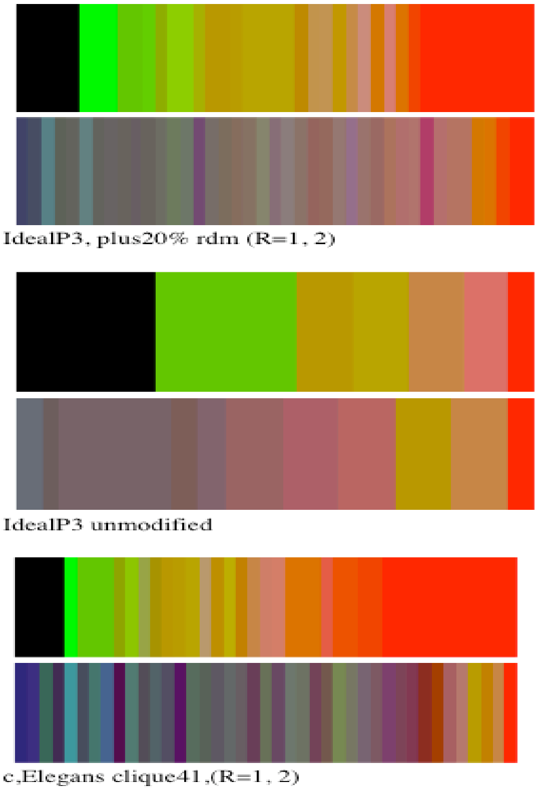

Fig 8. Middle pair: Colored spectra for the ideal P3-1 polygonal tiling for radii 1 and 2; top pair: 20% rewired edges using a uniform random selection; bottom pair: the color spectra at radii 1 and 2 for the C.elegans aggregate 41v*.

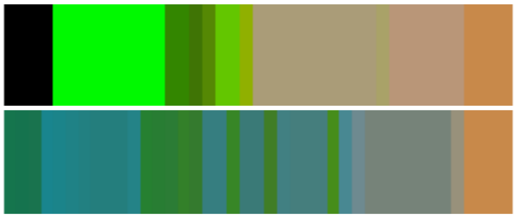

Fig 9. For comparison with the C.elegans aggregate in Fig 8, color spectra for a regular 6^3 triangular tiling of 41 nodes. based on radii 1 and 2. For the maximum radius, the spectrum is a blue green color.