# Comparing networks using their fine structure

Owen Macindoe* and Whitman Richards

CSAIL – 32-G585,
Massachusetts Institute of Technology,
Cambridge, MA 02139, USA
E-mail: owenm@mit.edu
E-mail: wrichards@mit.edu
*Corresponding author

**Abstract:** We introduce a novel technique for characterising networks using the structure of their sub-graphs, which we call the network's fine structure. To judge the similarities between networks we use the earth mover's distance between the distributions of features of their constituent sub-graphs. This technique is an abstraction of graph edit-distance. Given these similarity measures we explore their use in hierarchical clustering on several networks derived from a variety of sources including social interaction data.

**Keywords:** social network analysis; comparing networks; graph similarity; social computing.

**Reference** to this paper should be made as follows: Macindoe, O. and Richards, W. (2011) 'Comparing networks using their fine structure', *Int. J. Social Computing and Cyber-Physical Systems*, Vol. 1, No. 1, pp.79–97.

**Biographical notes:** Owen Macindoe is a graduate student of the Doctoral programme for Electrical Engineering and Computer Science at Massachusetts Institute of Technology. His research interests include social network analysis, graph comparison, automatic planning for collaborative tasks, game studies, and cognitive modelling. He is currently part of the Computer Science and Artificial Intelligence Laboratory's Learning in Intelligent Systems Group and the Singapore – MIT GAMBIT Game Lab.

Whitman Richards holds a BS degree and a PhD from MIT, with his thesis being a computational study of opponent process mechanisms underlying the Munsell Color Space. He is a Fellow of the Optical Society of America. He currently holds appointments in two departments at MIT (Brain and Cognitive Sciences, and Media Arts and Sciences) and resides at the Computer Science and Artificial Intelligence Lab. He has over 200 publications. One of his (edited) books, *Natural Computation*, lays out the groundwork for computational approaches to perception and manipulation. His present interests other than social networks include neural voting machines, constraints on collective choice and patterns of influence.

## 1 Introduction

Relevant to understanding a social network is whether its graphical form is similar to that of another network. For example, will a graph describing scientific collaborations be

similar to the graph of an e-mail network engaged in the development of Linux? Alternatively, we may have a theory of the graphical form of optimal organisational structure, and want to know how much an actual example deviates from this ideal. In both cases, we need to be able to judge graph similarity.

Consider two graphs $G$ and $H$ that are identical, except for a single edge absent in $H$. A natural way to think about judging their similarity would be to count the minimum number of changes that would have to be made to transform one graph into the other. This count is called the edit-distance and allows us to judge that a third graph $F$, missing two edges relative to $G$, is less similar to $G$ than $H$ is to $G$. Unfortunately, the problems with edit-distance are twofold. First, there are many possible kinds of edit operations, including edge rotation, edge addition and subtraction, and vertex addition and subtraction, and it is not clear how to weight these changes against one another. Additionally, to judge that an operation has in fact transformed one graph into the other involves solving the graph isomorphism problem, which has no known general polynomial time solution. It is clear that we will have to accept some level of approximation in any similarity measure for the sake of tractability.

We first briefly review previous attempts to overcome these problems and then present our own solution. We introduce a novel representation for graphs, which makes use of the distribution of structural features of their constituent sub-graphs, which we call a graph's fine structure. Using this representation we define graph similarity to be the earth mover's distance between these feature distributions and demonstrate that this abstraction yields sensible results under random graph permutation. We then go on to use this similarity measure to perform hierarchical clustering on a selection of networks, including social, neural, and semantic networks. Finally, we discuss the influence of a graph's generative process on graph similarity and discuss uses of our measure in investigating these processes.

## 2 Previous work

Some researchers have approached graph similarity using spectral analysis, where edit-distance is approximated by the difference in the spectrum of Eigenvalues between the laplacians of graph adjacency matrices (Peabody, 2002; McWherter, 2001). This was demonstrated in Peabody (2002) by cloning graphs, randomly permuting their copies, and showing that their spectral distance increases as a function of the amount of permutation. This technique has two weaknesses however, the first being the existence of isospectral graphs, which share Eigenvalues despite having quite different topological structure and therefore can erroneously be judged similar. The second is the difficulty of interpreting graph spectra as an abstraction of social phenomena. Ideally for the social network domain, we would like to design a similarity measure that judges graph similarity based on some set of features we suspect to be socially relevant.

Other related research includes $p^*$ models, graph kernels, and motif analysis. $p^*$ approaches to social network analysis typically attempt to fit the parameters of a class of exponential density functions, describing the probabilities of structures occurring within a graph, to empirically observed social graphs. These parameters can then be compared across graphs to judge their structural similarity (Anderson et al., 1999). Graph kernels are a broad class of functions that map graph features to points in high dimensional inner

product spaces, making them amenable to classification techniques such as SVMs (Shervashidze and Borgwardt, 2009; Borgwardt, 2007).

Motif analysis (Milo et al., 2002; Stoica and Prieur, 2009) computes the frequency of the occurrence of small sub-graphs, called motifs, and uses this analysis to judge the significance of the appearance of these motifs by comparison with their frequency in Erdős-Rényi random graphs. This work implicitly defines a similarity measure based on a comparison of motif frequencies. A key question for this approach is what is the right choice of motifs? If motifs are too large then the graph isomorphism problem arises again. If they are too small and numerous, then the high dimensionality of the feature space becomes unwieldy. What justifies a particular choice? Additionally, could some motifs be collapsed together into a single class of graphs, such as complete graphs or other special forms for the purposes of judging similarity? These considerations are part of the motivation for the *LBD* graph representation that we present in the next section.

## 3 The LBD representation

There are many possible choices for features that can abstractly represent the structure of a graph (Milo et al., 2002; Newman, 2003; Read and Wilson, 1998). For this work, we have chosen a triple of features, first introduced in Richards and Wormald (2009), that has some social relevance. These features are characterised as *leadership* (*L*), *bonding* (*B*), and *diversity* (*D*). We will use *LBD* triples to represent undirected graphs as points in *LBD* space.

### *3.1 Leadership*

Leadership, introduced in Freeman (1978), is a measure of the extent to which the edge connectivity of a graph is dominated by a single vertex. It is given by equation (1), in which $n$ is number of graph vertices and $d_i$ is the degree of vertex $i$. It is a normalised difference between the degree of the highest degree vertex and each other vertex in the graph. Leadership is maximal (i.e., 1) in a star graph (one vertex of degree $n - 1$ with all other vertices of degree 1) and zero for regular graphs with all vertices having the same degree (e.g., a complete graph or a ring). In a social network, a high leadership indicates that a small number of people are connected to a much larger proportion of others than the average group member, whereas a low leadership indicates that most people are equally connected.

$$L = \frac{\sum_{i=1}^{n} \left( d_{\max} - d_i \right)}{(n-2)(n-1)} \tag{1}$$

### *3.2 Bonding*

Bonding, given by equation (2), measures triadic closure in a graph. It is the ratio of length three paths in a graph to length two paths and is one of several measures called clustering coefficient in the literature (Wasserman and Faust, 1994). The motivation behind bonding is that this ratio measures the proportion of triadic closures that actually exist in a graph relative to the number that could exist, but are missing an edge. Bonding

is maximal (i.e., 1) for a complete graph, but zero for any graph with no triangle sub-graphs (e.g., trees or bipartite graphs). In a social network, a high bonding means that if two people are linked to a third person, then it is likely that they are also linked to one another. Where edges represent friendship for example, a high bonding means that if two people are mutually friends with a third person, then they are also likely to be friends with one another.

$$B = \frac{6 \times (\#\ \text{triangles})}{\#\ \text{length\_two\_paths}} \tag{2}$$

### *3.3 Diversity*

Diversity, given by equation (3), is a measure based on the number of edges in a graph whose end vertices are not adjacent, and hence are disjoint. We call such end vertices disjoint dipoles. The maximum number of disjoint dipoles for any graph of order n is the maximum number of four cycles in a graph of the same order. This maximal count is used as a normalising factor. The square root of the ratio scales the measure into a range similar to $L$ and $B$ (see Richards and Wormald, 2009, for details.) $D = 0$ for $n < 4$ and possible values lie in the range [0, 1]. Diversity is high in graphs which are not densely connected, such as bipartite graphs, but also in graphs where separate graph regions are joined by a relatively small number of bridging edges. In a social network, a high diversity indicates that separate communities exist, where people from one community have no direct ties with people in another, whereas a low diversity indicates that people are generally all connected to one another.

$$D = \sqrt{\frac{\#\ \text{disjoint\_dispoles}}{\left(\frac{n}{4}\left(\frac{n}{2}-1\right)\right)^2}} \tag{3}$$

Taken together, $L$, $B$, and $D$ summarise a graph's structure along three socially relevant dimensions. Plotting graphs in this space is a first step in determining which graphs are similar to one another. An abstract measure of the similarity of two graphs would then be the inverse Euclidean distance between two graphs in this feature space. We return to this idea in Section 6.

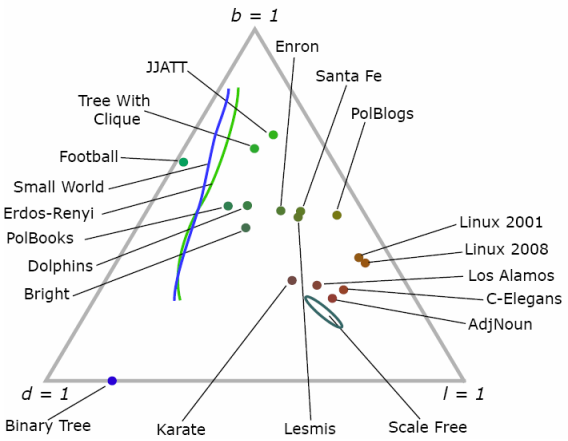
## 4 The *Lbd* simplex

Given the *LBD* scores of a graph we can plot its position in a feature space, with $L$, $B$, and $D$ scores being orthogonal axes. For ease of visualisation, however, we make use of a 2D slice of the 3D feature space. We first compute a graph's normalised $L$, $B$, and $D$ scores, which we call $l$, $b$, and $d$ respectively, by dividing each score by the sum of the three. This means a graph's *lbd* triple is a projection to a point on a two-simplex, which we call its simplex representation. A point in the centre of the simplex shows that the graph it represents is balanced along the three dimensions, whereas a point that is closer to a vertex of the simplex is dominated by a particular feature. See also Richards and

Macindoe (2010b) for an alternative visualisation that encodes *LBD* scores as colours in RGB space.

Figure 1 shows the *lbd* position of the networks analysed in this paper, as well as some graphs with well known structures. Points in the simplex are coloured according to their position in *lbd* space, with the red, green, and blue colour components corresponding to *l*, *b*, and *d* respectively. The two loci shown on the left of the simplex show the range of *lbd* scores that result from different parameter settings of Erdős-Rényi random graphs and Watts-Strogatz small world graphs with 300 vertices and density and rewiring parameters respectively in the range [0.2, 0.8]. The locus on the right shows the range of *lbd* scores for 300 vertex scale free graphs generated using the Barabasi-Albert preferential attachment model with the parameter setting for the number of edges added with each new node ranging from 2 to 8. The social networks analysed cover the upper right space in the simplex between these loci and notably do not fall in the same regions as these models. The majority of the networks analysed cover a restricted region in the simplex, yet it is possible to construct graphs whose *lbd* scores fall in other regions of the simplex (Richards and Macindoe, 2010a).

**Figure 1** Different networks have a wide range of LBD scores (see online version for colours)




Note: Here, to clarify, their positions are projected onto the (1, 1, 1) plane (i.e., the *lbd* simplex).

## 5 *LBD* distributions

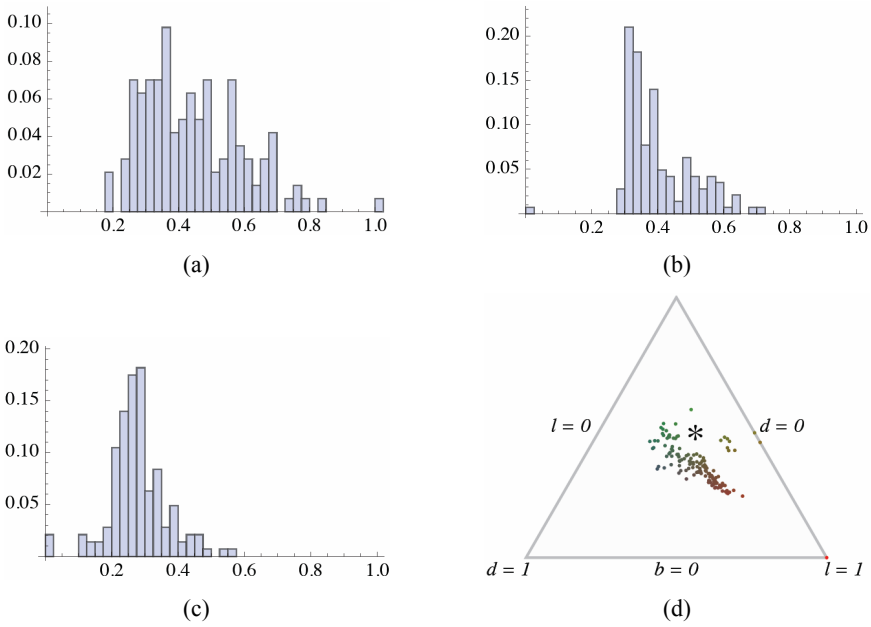
The *LBD* and *lbd* representations of a graph gives concise summaries of properties of the graph as a whole. But consider the case of a graph with multiple topologically distinct regions, an extreme example of which might be a series of cliques joined together in a chain by bridging edges. The cliques and their links would constitute a smaller scale structure, which we loosely call the graph's 'fine structure'. We would ideally like our representation to be fine grained enough to distinguish between this kind of graph and another graph without this fine structure that happens to map to the same *LBD* value or has the *LBD* scores in the same relative proportions, and hence the same *lbd* scores. More

generally, we would like a representation that reveals features of the fine structure of a graph and can answer such questions as whether the local sub-graphs centred on any given vertex in the graph are homogenous or heterogenous across the full graph. The graph described above is an example of a graph with heterogenous fine structure, whereas a ring is an example of a homogenous graph.

We represent the fine structure of a graph as the distribution of *LBD* values of its constituent sub-graphs. Specifically, a graphs' *LBD* distribution is a normalised histogram of the *LBD* scores of all the induced sub-graphs centred on each of its vertices. These distributions have a scale parameter, namely the radius of the sub-graphs, which controls the coarseness of the analysis. For example, to compute the radius 1 *LBD* distribution for a graph, we iterate over every vertex in the graph, computing an *LBD* score for the induced sub-graph formed by the vertex, its neighbours, and all the edges connecting them. Normalising the histogram counts by the size of the graph then yields a distribution over *LBD* scores. Note that as the radius of the *LBD* distribution approaches the diameter of the graph, the histogram will converge to a spike on the *LBD* score of the full graph, since in the limit each induced sub-graph will contain all of the graph's vertices and edges.

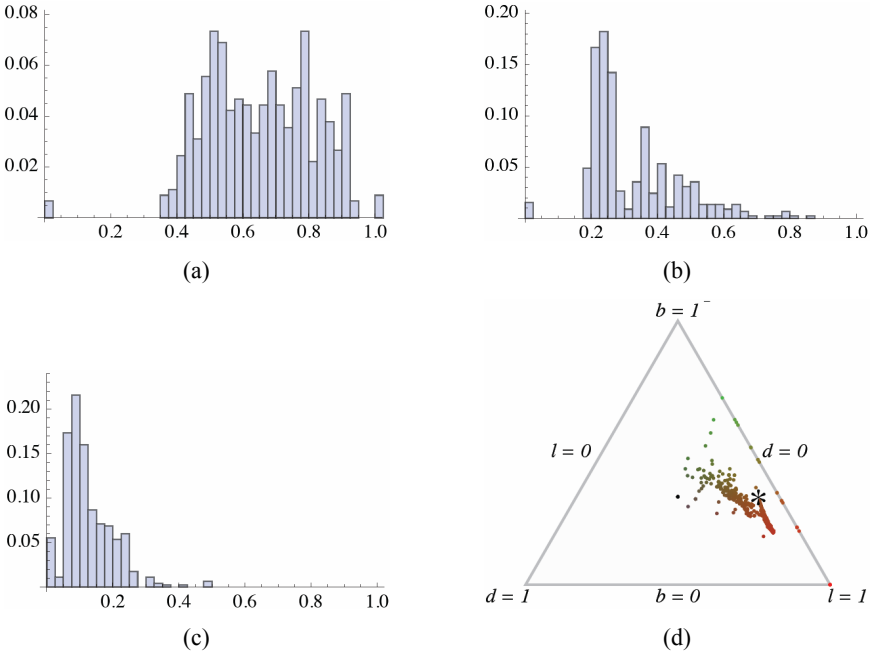
The *LBD* distribution can be thought of as an abstraction of the distribution of motifs produced by motif analysis. Any given motif has an associated *LBD* value, but some motifs may map to the same value; for instance all star graphs, regardless of size, map to $L = 1$, $B = 0$, $D = 0$. The LBD distribution then is akin to a motif distribution which generalises across classes of motif based on their *LBD* score.

**Figure 2** *LBD* distributions and simplex for the Enron network at radius 2, (a) leadership (b) bonding (c) diversity (d) *lbd* simplex (see online version for colours)




Notes: The asterisk indicates the *lbd* location for the full graph (i.e., radius is now the diameter of the graph). The histograms show the frequencies of the parameters, given on the abscissa.

**Figure 3** *LBD* distributions and simplex for the Linux-2008 network at radius 2, (a) leadership (b) bonding (c) diversity (d) *lbd* simplex (see online version for colours)

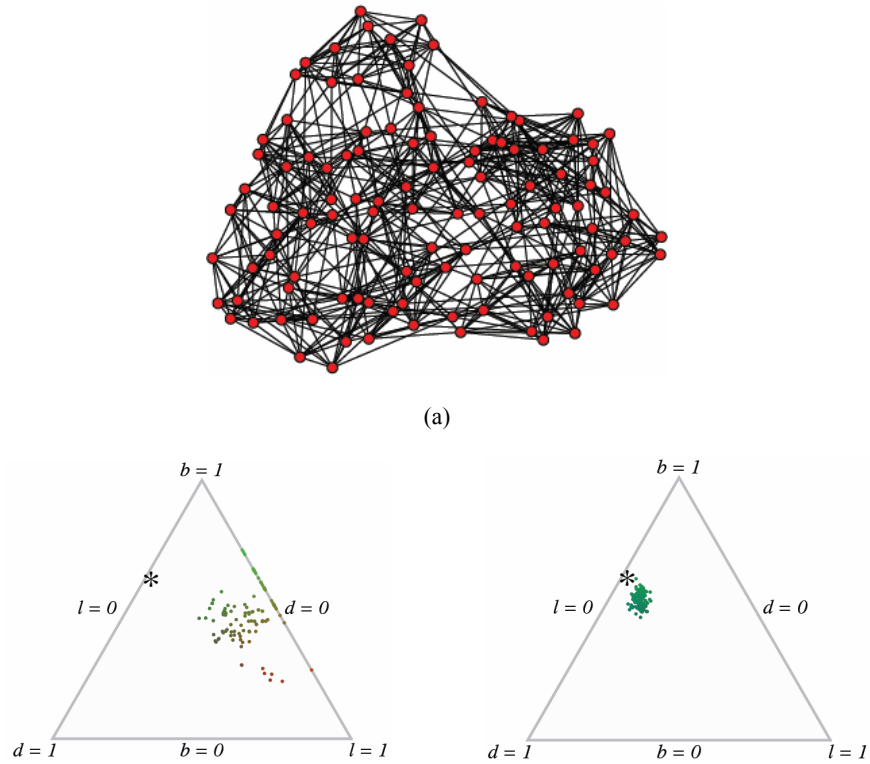



Notes: The asterisk indicates the *lbd* location for the full graph (i.e., radius is now the diameter of the graph). The histograms show the frequencies of the parameters, given on the abscissa.

Figures 2 and 3 show the radius 2 *LBD* distributions for two e-mail exchange networks. The first is extracted from the Enron e-mail dataset collected a part of the CALO project (Cohen, 2009). Each vertex is an e-mail address in the dataset and an edge links two vertices if the e-mail addresses both sent at least one e-mail to each other. E-mail addresses that only sent and never received or vice-versa were not included. The second comes from an analysis of Linux kernel mailing list traffic in January of 2008 compiled by Gnawali (2009). Here, each vertex is again an e-mail address, with some e-mail aliases being collapsed into a single vertex. An edge again indicates that at least one e-mail was exchanged each way between the two addresses. From the distributions in the two figures we can see that the sub-graphs comprising the Linux graph tend to have higher leadership scores, but lower bonding scores than those in the Enron graph. For Linux, this suggests that locally, people tend to communicate with highly connected individuals rather than directly with others in their neighbourhood. For Enron, the marginally higher bonding suggests more direct communication between people in local neighbourhoods and the lower leadership indicates there are fewer people who are involved in a disproportionately large number of different e-mail conversations than is the case with Linux. The higher diversity score in the Enron graph suggest a somewhat more fractured local graph structure, which together with the higher bonding is indicative of more groups of people who largely do not correspond with each other, being joined by a small number of common members. This makes sense for an organisation such as Enron where team members might e-mail one another and managers or team leaders

serve as communication bridges between teams. It is interesting that the full graph *LBD* score for the Linux graph is close to its radius 2 cloud of points in the simplex, whereas this is not the case for Enron. This demonstrates how in some cases the fine structure of a graph can be quite different from the structural features of the graph considered as a whole.

**Figure 4** Visualisation and *lbd* simplexes for the football network at radii 1 and 2, (a) the football network (Girvan and Newman, 2002), note the clustering of teams into local competitions (b) radius 1 (c) radius 2 (see online version for colours)



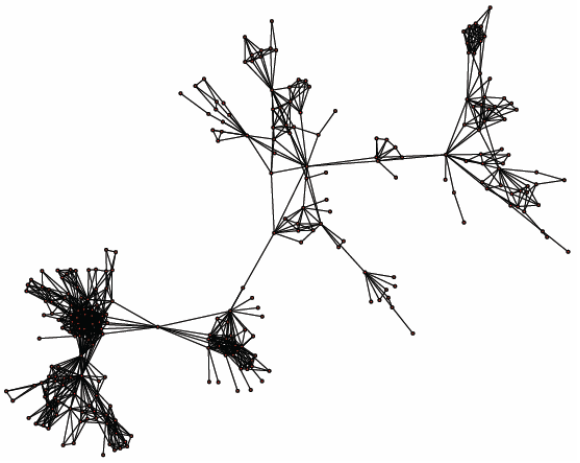

(a) (b) (c)

LBD distributions can look dramatically different across different radii. Figure 4(a) shows a highly structured graph of football matches between division IA colleges in Fall of 2000 compiled by Girvan and Newman (2002), in which each vertex is a team and each edge is a match. The general structure is that local teams play one another, forming small bonded sub-graphs, and then their winners play one another, linking the sub-graphs. Figures 4(b) and (c) show the distribution of *lbd* values at radius 1 and radius 2. At radius 1, we can see that a large proportion of the graph is composed of sub-graphs with one or two vertices whose degree is higher than the rest of the vertices in the sub-graph. These vertices are division winners and their influence can seen in the mid to high range leadership values in the simplex. As is typical of radius 1 sub-graphs, diversity scores tend to be low. This tells us that when we look at just the sub-graph of a team and the teams that they have played against, there are one or two teams that have

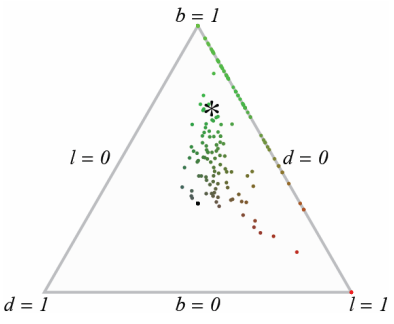
played more games and that most teams have played games against opponents within their own local competition. At radius 2, there is a dramatic shift. Since the graph has a low diameter, radius 2 neighbourhoods include most of the graph, leading to a convergence in *lbd* scores. Leadership scores become much lower, because now most sub-graphs include most division winners which compete with one another in degree. Diversity also rises as different divisions are linked by the winners of those divisions playing one another. At higher radii, the point cloud converges towards the asterisk, which shows the full graph *lbd* score.

By way of contrast, Figure 5 shows associations between actors involved in terrorist attacks in Southeast Asia drawn from the John Jay and Artis Transnational Terrorism Database (Atran et al., 2008). Note that both the graph and simplex visualisations show a range of different structures in the network's local neighbourhoods, ranging from tightly connected groups to loose associations of individuals, but also that the distribution of *lbd* scores changes more slowly across radii than the Football network due to the large diameter of the network.

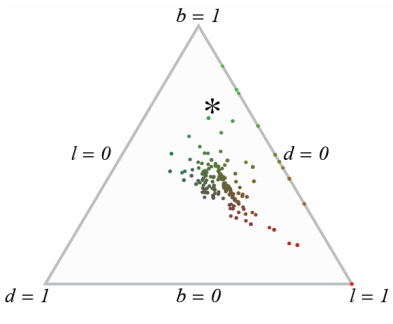
**Figure 5** Visualisation and *lbd* simplexes for the JJATT network at radii 1 and 2, (a) the JJATT network (Atran et al., 2008) contains a diverse spread of sub-networks (b) radius 1 (c) radius 2 (see online version for colours)



(a)




(b)




(c)

## 6 Comparing graph fine structure

Since the *LBD* distribution of a graph summarises its fine structure we can compare the *LBD* distributions of two graphs to judge their similarity. In performing this comparison, there are some choices and tradeoffs to be made. The first is what radius to consider for the distributions. For much social network analysis, researchers are interested in ego-centric sub-graphs within a social network, which corresponds to a radius 1 analysis, or perhaps radius 2 if they are interested in an analysis of the structure of the sub-graphs including friends of friends. From our experiments, the most interesting results come from analysis at these two radii, particularly radius 2, at which sub-graphs become large enough for diversity to be a significant factor.

An issue which was not mentioned in Section 5 is whether or not to make the *LBD* space discrete when computing distributions. *LBD* distributions were derived from counts of the occurrences of real valued *LBD* scores for sub-graphs. However, for the purposes of ease of comparison we may wish to bin *LBD* values within discretised regions. The choice of the granularity of this discretisation will impact any comparison, since coarser discretisations may place distinct points in the same bin. We chose a compromise between abstraction and fidelity by discretising *LBD* space into 0.2 unit length cubes with the result that some graphs may be judged more similar than in the non-discretised case. Our results suggest however that the discretisation process does not introduce an unreasonable amount of noise.

Another concern relates to the question of what kind of comparison of fine structure we want to make. Our construction of *LBD* distributions weights each discretised *LBD* region's contribution in the representation by the proportion of sub-graphs that fall into that region. An alternative construction would be simply a vector of *LBD* values occurring in the graph. The distinction here is that in the former representation proportion is important, whereas in the latter mere presence is important. Consider for instance the case where two graphs were being compared and our criterion for similarity were whether one is a sub-graph of the other, larger graph. In this case, perhaps the presence-based representation may be more appropriate for comparison than our proportional representation. This consideration makes clear that in comparing *LBD* distributions we are comparing the relative proportions of the features of the graphs' fine structure. An upshot of this approach is that because we normalise the distribution, a comparison between two graphs of different sizes is possible, whereas in a presence-based representation this would add complications.

We begin our fine structure comparison by choosing a sub-graph radius, $r$, and computing histograms, with bin sizes of 0.2, of the *LBD* scores of the radius $r$ induced sub-graphs in each graph. We then normalise the counts of the histogram bins by dividing by the number of vertices in each graph, yielding two *LBD* distributions. We compute the earth mover's distance (Pele and Werman, 2008, 2009; Rubner et al., 1998) between these two distributions using Euclidean distance as the ground distance. Finally, we normalise by the maximum distance in the discretised space and subtract the result from 1 to yield a similarity measure in the range [0, 1]. To demonstrate that this similarity measure produces intuitively plausible results, we followed the example of Peabody (2002) and computed the similarity of a variety of graphs to permutations of themselves. We used this technique on a set of graphs from a number of sources and modelling a wide variety of phenomena, from social networks and e-mail traffic to football match-ups and neural networks. Table 1 gives an overview of the graphs included in the analysis,

showing the number of vertices (| *V* |), edges (| *E* |), edge probability [*P*(*E*)], characteristic path length (*CPL*), diameter, and full graph *LBD* scores. Three artificial graphs not used in the analysis are included in the table for the sake of comparison; a binary tree, a small tree with a clique attached to one leaf, and an Erdős-Rényi random graph with a density parameter of 0.09. Where graphs originally contained directed or weighted edges, these were converted to unweighted and undirected edges, and this loss of structure must be kept in mind when interpreting the results of our analysis. To produce the permutations we chose a percentage of noise and randomly permuted that proportion of edges in the original graph. The similarity as a function of permutation averaged over ten trials for a variety of graphs is plotted in Figure 6, which demonstrates, as hoped, that our similarity measure judges graphs to be less similar to their permutations as the degree of permutation increases. As a twist on this result we performed the same process on an Erdős-Rényi random graph with 115 vertices and edge probability 0.09. This is the top line in the plot, almost coincident with the top of the figure. The consistent high similarity score shows that permuting a random graph does not necessarily make it dissimilar to itself. This is because the construction of Erdős-Rényi random graphs with such an edge probability leads them to have characteristic fine structure properties, namely low leadership, low bonding, and high diversity. The effect of permuting a random graph is to transform it into another random graph with the same parameters. Note also that there is a lower bound for each graph on self-dissimilarity caused by permutation, which is related to how close the original graph's *LBD* distribution is to the region typical of Erdős-Rényi random graphs.

**Table 1** Vertices, edges, edge density, characteristic path length, diameter, and *lbd* scores for the analysed graphs

| *Graph* | \| *V* \| | \| E \| | *P*(*E*) | *Diameter* | *CPL* |
|---|---|---|---|---|---|
| Los Alamos | 30 | 78 | 0.1793 | 4 | 2.0598 |
| Karate | 34 | 78 | 0.1390 | 5 | 2.4082 |
| Dolphins | 62 | 159 | 0.0841 | 8 | 3.3570 |
| Enron | 143 | 623 | 0.0614 | 8 | 2.9670 |
| Santa Fe | 116 | 174 | 0.0261 | 15 | 6.6576 |
| JJATT | 263 | 998 | 0.0290 | 13 | 5.8750 |
| Linux 2001 | 302 | 749 | 0.0165 | 7 | 3.1614 |
| Linux 2008 | 447 | 2,122 | 0.0213 | 6 | 2.7919 |
| Bright | 54 | 175 | 0.1223 | 5 | 2.5947 |
| Lesmis | 77 | 254 | 0.0868 | 5 | 2.6411 |
| PolBooks | 105 | 441 | 0.0808 | 7 | 3.0788 |
| Adj-Noun | 112 | 425 | 0.0684 | 5 | 2.5356 |
| Football | 115 | 613 | 0.0935 | 4 | 2.5082 |
| C-Elegans | 297 | 2148 | 0.0489 | 5 | 2.4553 |
| PolBlogs | 1,490 | 16,715 | 0.0151 | 8* | 2.7375* |
| Binary tree | 127 | 126 | 0.0157 | 12 | 8.3510 |
| Tree with clique | 62 | 496 | 0.2623 | 10 | 5.2512 |
| Erdős-Rényi | 115 | 598 | 0.912 | 4 | 2.2632 |

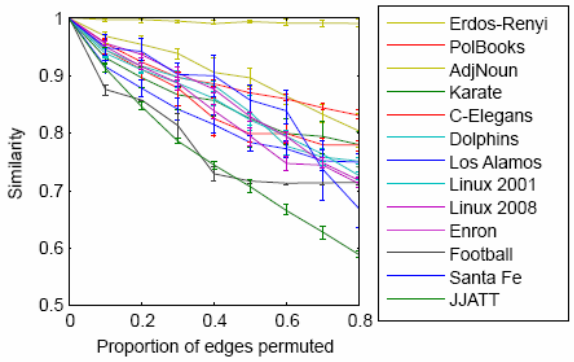
Note: The asterisk indicates that the PolBlogs graph is not connected and the reported values are for its largest component.

**Table 1** Vertices, edges, edge density, characteristic path length, diameter, and *lbd* scores for the analysed graphs (continued)

| *Graph* | *L* | *B* | *D* | *Type* | *Source* |
|---|---|---|---|---|---|
| Los Alamos | 0.6946 | 0.3683 | 0.2923 | Co-authorship | Palla et al. (2005) |
| Karate | 0.3996 | 0.2557 | 0.2402 | Social | Zachary (1977) |
| Dolphins | 0.1164 | 0.3088 | 0.1959 | Social | Lusseau et al. (2003) |
| Enron | 0.2377 | 0.3591 | 0.1455 | E-mail | Cohen (2009) |
| Santa Fe | 0.1681 | 0.2200 | 0.0683 | Co-authorship | Girvan and Newman (2002) |
| JJATT | 0.1362 | 0.4905 | 0.0744 | Social | Atran et al. (2008) |
| Linux 2001 | 0.2510 | 0.1534 | 0.0333 | E-mail | Gnawali (2009) |
| Linux 2008 | 0.3435 | 0.1929 | 0.0393 | E-mail | Gnawali (2009) |
| Bright | 0.2257 | 0.3770 | 0.2634 | Semantic | Palla et al. (2005) |
| Lesmis | 0.3972 | 0.4989 | 0.1755 | Literature | Knuth (1993) |
| PolBooks | 0.1627 | 0.3484 | 0.1877 | Economic | Krebs (2003) |
| Adj-Noun | 0.3799 | 0.1569 | 0.1320 | Semantic | Newman (2006) |
| Football | 0.0120 | 0.4072 | 0.2355 | Sports | Girvan and Newman (2002) |
| C-Elegans | 0.4066 | 0.1807 | 0.1106 | Neural | White et al. (1986) |
| PolBlogs | 0.2210 | 0.2260 | 0.0327 | Citation | Adamic and Glance (2005) |
| Binary tree | 0.0082 | 0.0000 | 0.04355 | Artificial | Macindoe (2010) |
| Tree with clique | 0.2541 | 0.9945 | 0.2570 | Artificial | Macindoe (2010) |
| Erdős-Rényi | 0.0768 | 0.0853 | 0.2110 | Artificial | Macindoe (2010) |

Note: The asterisk indicates that the PolBlogs graph is not connected and the reported values are for its largest component.

**Figure 6** Radius 2 self-similarity under random edge permutation (see online version for colours)

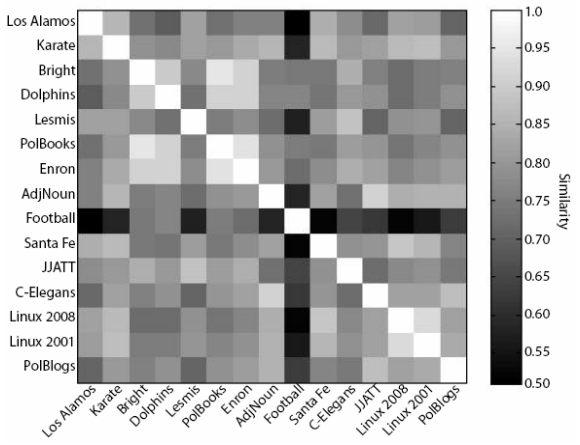



## 7 Clustering graphs

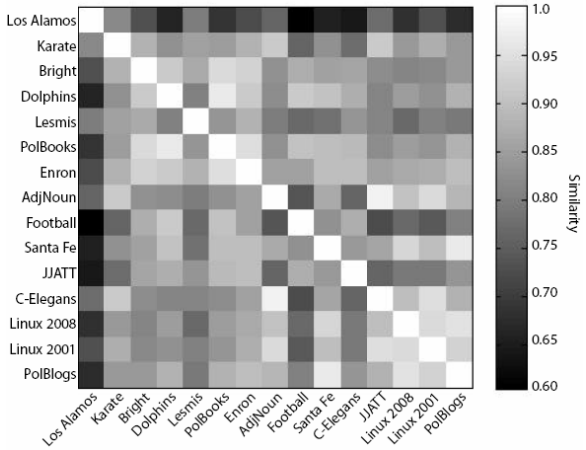
Armed with a method for judging graph similarity by fine structure features, we use it to find classes of graph that have these features in common. Using a hierarchical clustering approach we can take a set of graphs and find clusters of graphs that are similar to one

another but dissimilar to graphs outside their cluster. There are many choices of clustering algorithm available, so we opted for the generality and simplicity using average-link hierarchical clustering following the method in Dunham (2002). In this agglomerative approach to clustering we compute the pairwise similarities of all the graphs in the set to be clustered. Initially, each graph is in its own cluster. At each step we then merge the two clusters for whom the mean similarity is highest, resulting in a hierarchy of graph clusters. Since there is no gold standard of graph groupings against which to judge the outcome of the clustering, this should be viewed as an exploratory analysis.

**Figure 7** Radius 2 and full graph similarities, (a) radius 2 graph similarities (b) full graph similarities




(a)



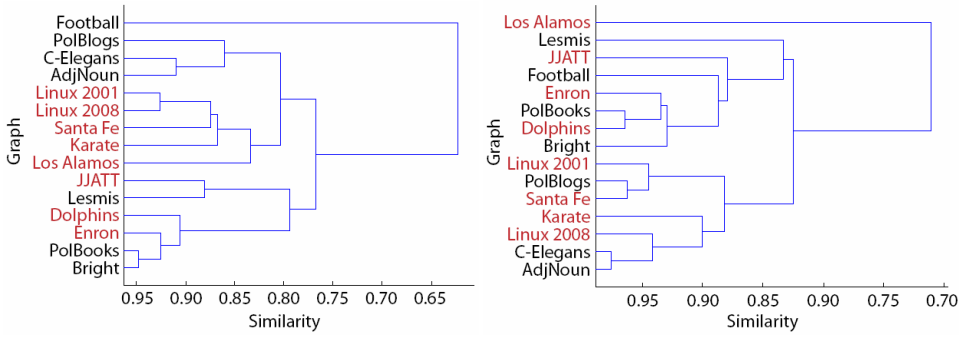

(b)

We performed our clustering analysis on the same set of graphs listed in Table 1. Figure 7(a) shows the pairwise similarity between each graph in the set computed with a radius of 2. By contrast, Figure 7(b) shows the similarity between the graphs judged by

the inverse of distance between their full-graph *LBD* scores. Contrasting these results, it is clear there is a qualitative difference between similarity judged at the full graph level and similarity judged at the fine structure level. This is particularly visible in the distinctive dissimilarity of the football graph from other graphs in the set, judged by the fine structure analysis which has discovered the structural regularities in the graph that result from the generative process of match-making that forms it and gives it the locally homogenous structure that we saw in Section 5. The general conclusion we can draw from this is that two graphs can have a similar global structure, judged by their full graph LBD score, and yet have quite dissimilar fine structures.

Figures 8(a) and (b) show dendrograms for the results of the clustering using the radius 2 and full-graph similarity respectively. Horizontal lines represents clusters, with lines joining at a given similarity, shown on the horizontal axis, indicating that two clusters were chosen to be merged at that similarity threshold. The names of graphs derived from social data, such as e-mail correspondence or co-authorship are shown in red. Again, a key point is that the results are different, indicating that similarity in fine structure and full graph structure are not equivalent.

**Figure 8** Hierarchical clustering dendrograms based on radius 2 and full graph similarity, (a) radius 2 fine structure dendrogram (b) full graph dendrogram (see online version for colours)




(a) (b)

Note: Graphs with red names are derived from social data.

Looking at the clusters formed by the fine structure analysis, note that they often contain a mix of different kinds of graphs, for instance Bright, a semantic network, and PolBooks, a graph of book co-purchases, have the most similar fine structures. Other clusters are more homogenous, for instance the two Linux graphs are placed in the same initial cluster, which suggests that there is consistency in the way that e-mail correspondence on the Linux mailing list is structured over time. The Linux graphs in turn form part of a larger cluster that contains the majority of the social graphs, yet interestingly does not contain Enron, the other e-mail correspondence graph in the dataset. AdjNoun, a semantic network, and C-Elegans, a neural network, are the only two graphs that are judged as being more similar to each other than to any other graphs in the dataset in both the full graph and fine structure analyses. This fine structure similarity judgement stems from the fact that in both cases the LBD distributions of the radius 2 sub-graphs of both these graphs balance bonding and diversity against one another whilst having a high-skewing spread of leadership scores.

The dissimilarity of the football graph from all other graphs, judged by its fine structure, is again due to a combination of its small radius, which leads to its radius 2 sub-graphs being relatively homogenous, and the fact that there is low variation in the degree of its vertices, which leads to low leadership scores that are uncommon in other graphs such as social networks, which tend to contain more variation in connectivity. These considerations lead it to be placed in a cluster by itself in the fine structure analysis, whereas the full graph clustering does not respond to its unusually homogenous fine structure.

Neither the full-graph nor the fine structure similarity measure judges the collaboration networks Santa Fe and Los Alamos to be particularly similar. In the case of fine structure, this is most likely because the small number of vertices in the Los Alamos graph makes its distribution much more sparse along the leadership axis than the Santa Fe graph, even though the bonding and diversity scores fall in a similar range. At the full graph level, the differences are even more pronounced, with the Los Alamos graph having a much higher leadership and bonding than Santa Fe. Together these suggest that the idiosyncratic characteristics of a particular group of collaborators are more crucial to the formation of a graph's structure at both a macro level and in its fine structure than the mere fact that the graph represents people collaborating on papers as opposed so some other activity such as corresponding via e-mail.

It is also interesting to note that both analyses make very similar judgements about the higher level clustering of the graphs. Both methods judge that there is one hierarchical cluster containing JJATT, Dolphins, Enron, PolBooks, Bright, and Lesmis and another containing AdjNoun, C-Elegans, PolBlogs, Karate, Santa Fe, and the two Linux graphs, with some disagreement about the placement of Football and Los Alamos, which are in a sense exceptional due to either their homogenous structure or small size. At the fine structure level these cluster distinctions seem to be related to the tightness of the spread along the leadership dimension, but at level of similarity at which these two clusters are finally merged the intra-cluster similarities are themselves quite low, making a general characterisation of the distinct clusters hard.

Finally, note that in the fine structure clustering the majority of the graphs drawn from social data are placed together in one homogenous cluster. The excluded graphs are JJATT, which exhibits unusually high *B* scores in its sub-graphs, Dolphins, which is from non-human social data, and the Enron e-mail graph. By contrast, the clustering based on full graph *LBD* scores produces clusters that are very mixed with respect to the source of their graph data.

## 8 Discussion

In the previous section, we identified clusters of graphs with similar features in their fine structure. The natural question then is how does this common structure arise? One of the hopes of network analysis is that studying their structure may give clues as to the generative processes that underlie their formation. One might ask whether the structural features arising from some proposed model of network growth leads to the kind of structure seen in empirical data on networks. For example, which networks might be the result of the preferential attachment of new vertices to well connected old ones, as in Barabasi-Albert scale free networks, or which networks resemble the small world constructions of Watts and Strogatz?

In Figure 1, we showed the loci of full graph *lbd* scores displayed for a range of parameters for several well studied models, namely Erdős-Rényi random graphs, Watt-Strogatz small world graphs, and Barabasi-Albert scale free graphs. These loci characterise the range of full graph structural features that these models generate when used to produce graphs of roughly the same order as the empirical networks analysed. The fact that the *lbd* scores of empirically observed networks do not lie within any of these loci is evidence against any of these models being adequate characterisations of the generative processes that produced them. Although not shown here, this negative result is also supported by clustering results for these models using the fine structure similarity measure presented in Section 6, which place graphs generated by the models in clusters that are separate from the empirically observed networks.

An unachieved objective was to characterise typical *LBD* distributions produced by well studied generative processes. As a first step towards this kind of analysis we can point to our investigation of the self-similarity of permuted Erdős-Rényi random graphs in Section 6 as an example of evidence for the existence of homogenous fine structure across different instantiations of a given model of network formation. Recall that in permuting the edges of a random graph the effect was to transform one Erdős-Rényi random graph into another instance of a random graph with the same parameters and that all of these instances were judged similar by our fine structure comparison technique. This empirical observation suggests that it is a property of the process that generates Erdős-Rényi random graphs that causes their fine structure to tend to be similar, but more analytic work is needed to prove this in the general case. Characterising the range of *LBD* distributions that other models generate is an important challenge, with only baby steps currently in place (Richards and Macindoe, 2010a), but initial investigations strongly suggest that random graph, scale free, and small world network growth models do not adequately characterise the empirically observed fine structure of social networks of the scale presented in this paper.

On a more positive note, the fine structure analysis of some pairs of networks in Section 6 revealed some compelling pairwise similarities, for instance the Linux correspondence networks are very similar both in their full graph and fine structure, despite being drawn from data generated years apart. This suggests the existence of a consistent generative process responsible for this homogenous structure, which could in principle be modelled. However, we also have evidence that the generative process that produces graphs representing the same phenomena, for instance e-mail correspondence graphs, can be quite idiosyncratic. One might expect that if the Enron and Linux correspondence graphs were generated by a similar process, then their fine structure should be similar too, but in fact neither their fine structure nor their full graph structure is similar, which suggests that dissimilarities in the organisational structures of Enron and the Linux kernel developers are more crucial factors in the formation of the graphs than the mere fact that the graphs represent e-mail correspondence. As mentioned in the previous section we can draw similar conclusions for collaboration graphs.

Our conclusion is that the specific conditions under which the phenomena that a graph models take place can be more crucial for its fine structure characteristics than the general class of phenomena that the graph represents. A core challenge for further research then is to characterise these conditions and the generative processes to which they give rise. Our fine structure analysis technique is a key a tool for judging the

plausibility of a proposed generative process by providing a method for judging the similarity between the fine structure of an empirically observed graph and graphs produced by a proposed model. Our technique can also help identify networks that may have common generative processes by highlighting networks that have high fine structure similarity at a range of local neighbourhood radii. Identifying the commonalities between these networks may then help develop models of generative processes that explain these structural similarities.

## 9 Conclusions

The key contribution of this paper is the introduction of a method for comparing the fine structure of graphs based on socially relevant features. The method makes use of the distribution of structural features of the sub-graphs that comprise the local neighbourhoods within the network at a given scale of granularity, which we called the network's *LBD* distribution at that granularity. These features summarise structural characteristics that are particularly relevant for social networks, yet are general enough to be relevant for large classes of graphs. We demonstrated that the choice of granularity, controlled by the radius of the sub-graphs for which an *LBD* distribution is computed, can have a strong effect on the shapes of distributions and by extension the similarity measures computed from them.

We demonstrated that our method produces intuitive results when comparing graphs against permutations of themselves and then used the measure to cluster a diverse set of graphs. We contrasted our clustering with that produced by a method that judged similarity based simply off the *LBD* score for a full graph and showed that the fine structure-based clustering gave a better agreement in some cases with our intuitions, for instance judging two graphs of e-mail correspondence from the Linux kernel mailing list to be similar in contrast with the full graph *LBD* clustering.

We noted for the set of graphs we were analysing that their fine structural similarity did not seem to be dependent upon the phenomena that the graphs were modelling. This led us to conclude that idiosyncratic features of organisations were likely to have more influence on a graph's fine structure than broad commonalities between people's e-mail correspondence or collaborative research behaviour. Furthermore our analysis showed that graphs can be judged similar by their full graph structure and yet dissimilar by their fine graph structure, emphasising the importance of choosing the granularity of analysis at which a similarity judgement is to be made.

Our technique is a useful tool both for comparing empirical graphs and for comparing the fine structure of graphs produced by a proposed generative process to the empirically observed graphs that they are seeking to explain.

## Acknowledgements

The authors would like to thank Omprakash Gnawali for providing the Linux e-mail data and also the AFOSR for support under a MURI grant A9550-05-1-0321.

## References

Adamic, L.A. and Glance, N. (2005) 'The political blogosphere and the 2004 US election', *Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem*.

Anderson, C.J., Wasserman, S. and Crouch, B. (1999) 'A p* primer: logit models for social networks', *Social Networks*, Vol. 21, pp.37–66.

Atran, S., Bennett, S., Fatica, A., Magouirk, J., Noricks, D., Sageman, M. and Wright, D. (2008) *John Jay & ARTIS Transnational Terrorism (JJATT) Dataset*, available at http://doitapps.jjay.cuny.edu/jjatt/.

Borgwardt, K.M. (2007) 'Graph kernels', PhD dissertation, Ludwig Maximilians University, Munich.

Cohen, W.W. (2009) *Enron Email Dataset*, available at http://www.cs.cmu.edu/enron/.

Dunham, M.H. (2002) *Data Mining: Introductory and Advanced Topics*, Prentice Hall, New Jersey.

Freeman, L.C. (1978) 'Centrality in social networks: conceptual clarification', *Social Networks*, Vol. 1, pp.215–239.

Girvan, M. and Newman, M.E.J. (2002) 'Community structure in social and biological networks', *Proceedings of the National Academy of Sciences*, Vol. 99, No. 12, pp.7821–7826.

Gnawali, O.D. (2009) 'Linux kernel email communication networks from January 2001 and 2008', *Personal Communication*.

Knuth, D.E. (1993) *The Stanford GraphBase: A Platform for Combinatorial Computing*, Addison-Wesley, Reading.

Krebs, V. (2003) *Books about US Politics Dataset (unpublished)*, available at http://www.orgnet.com/.

Lusseau, D., Schneider, K., Boisseau, O.J., Haase, P., Slooten, E. and Dawson, S.M. (2003) 'The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations', *Behavioral Ecology and Sociobiology*, Vol. 54, pp.396–405.

Macindoe, O. (2010) 'Investigating the fine grained structure of networks', Masters thesis, Massachusetts Institute of Technology, Cambridge.

McWherter, D. (2001) 'Approximate variations of graph matching and applications', Masters thesis, Drexel University, Philadelphia.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U. (2002) 'Network motifs: simple building blocks of complex networks', *Science*, Vol. 298, pp.824–827.

Newman, M.E.J. (2003) 'The structure and function of complex networks', *SIAM Review*, Vol. 45, pp.167–256.

Newman, M.E.J. (2006) 'Finding community structure in networks using the eigenvectors of matrices', *Physics Review E*, Vol. 74.

Palla, G., Derényi, I., Farkas, I. and Vicsek, T. (2005) 'Uncovering the overlapping community structure of complex networks in nature and society', *Nature*, Vol. 435, pp.814–818.

Peabody, M. (2002) 'Finding groups of graphs in databases', Masters thesis, Drexel University, Philadelphia.

Pele, O. and Werman, M. (2008) 'A linear time histogram metric for improved sift matching', *Proceedings of the European Conference on Computer Vision*.

Pele, O. and Werman, M. (2009) 'Fast and robust earth mover's distances', *Proceedings of the International Conference on Computer Vision*.

Read, R. and Wilson, R. (1998) *An Atlas of Graphs*, Oxford Press.

Richards, W. and Macindoe, O. (2010a) 'Characteristics of small networks', MIT-CSAIL, Tech. Rep. 033.

Richards, W. and Macindoe, O. (2010b) 'Decomposing social networks', *Proceedings of the Second IEEE Conference on Social Computing*.

Richards, W. and Wormald, N. (2009) 'Representing small group evolution', *Proceedings of the IEEE Conference on Social Computing*, p.232.

Rubner, Y., Tomasi, C. and Guibas, L.J. (1998) 'A metric for distributions with applications to image databases', *Proceedings of the International Conference on Computer Vision*.

Shervashidze, N. and Borgwardt, K.M. (2009) 'Fast subtree kernels on graphs', *Advances in Neural Information Processing Systems 22: Proceedings of the 2009 Conference*, Vancouver, Canada, pp.1660–1668.

Stoica, A. and Prieur, C. (2009) 'Structure of neighbourhoods in a large social network', *Proceedings of the IEEE Conference on Social Computing*.

Wasserman, S. and Faust, K. (1994) *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge.

White, J.G., Southgate, E., Thompson, J.N. and Brenner, S. (1986) 'The structure of the nervous system of the nematode c. elegans', *Philosophical Transactions of the Royal Society*, Vol. 314, pp.1–340.

Zachary, W.W. (1977) 'An information flow model for conflict and fission in small groups', *Journal of Anthropological Research*, Vol. 33, pp.452–473.