# Graph Comparison Using Fine Structure Analysis

Owen Macindoe
CSAIL - 32-G585
Mass. Inst. of Technology
Cambridge, MA, 02319
owenm@mit.edu

Whitman Richards
CSAIL - 32-364
Mass. Inst. of Technology
Cambridge, MA, 02319
wrichards@mit.edu

***Abstract*—We introduce a novel technique for comparing graphs using the structure of their subgraphs, which we call a graph's fine structure. Our technique compares graphs using the earth mover's distance between the distributions of summarizing features of their constituent subgraphs. We demonstrate the use of this technique as an abstraction of graph edit-distance and show its use in hierarchical clustering on several graphs derived from a variety of sources including social interaction data.**

## I. 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 email network engaged in the development of Linux? Alternatively, we may have a theory of the graphical form of optimal organizational 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 A and B that are identical, except for a single edge absent in B. A natural way to think about judging their similarity would be to count the 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 C, missing two edges relative to A, is less similar to A than B is to A. 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's 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 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 subgraphs, 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 graphs, 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.

## II. 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 [1], [2]. This was demonstrated in [1] 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 [3]. 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 [4], [5].

Motif analysis [6] computes the frequency of the occurrence of small subgraphs, 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 and the graph isomorphism problem arises again. If they are too small and numerous then the graph's representation as motif frequency counts becomes unwieldily high dimensional. 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.

## III. The LBD Representation

There are many possible choices for features that can abstractly represent the structure of a graph [6]–[8]. For this work we have chosen a triple of features that has some social relevance, first introduced in [9]. These features are characterized as *leadership* (L), *bonding* (B), and *diversity* (D). We will use LBD triples to represent undirected graphs as points in LBD space. In this section we review these features and present examples of L, B, and D values computed for various graphs.

### *Leadership*

Leadership, introduced in [10], 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 the mean 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.

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

### *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 [11]. 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 subgraphs (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 likely to be friends with one another.

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

### *Diversity*

Diversity, given by equation (3), is a measure based on the number of edges that share no common end points, and hence are disjoint. A normalization is imposed by the maximal count, which occurs for the complete bipartite graph. The square root of the ratio scales the measure into a range similar to L and B (see [9] 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.

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

Taken together, L, B, and D summarize a graph along three socially relevant dimensions. Plotting graphs in this space is a first step in determining which graphs are similar to one another. Figure 1 shows the position of the graphs analyzed in this paper, as well as some graphs with well known structures, using a transformation of the 3D LBD space into the 2D (1,1,1) plane. This is done by normalizing the L, B, and D scores for a graph by the sum of these scores, yielding the normalized scores $l$, $b$, and $d$ (in lower case to distinguish them from the unnormalized scores), which are then plotted in the simplex, showing the relative magnitudes of these features. Throughout this paper we make use of simplex visualizations like this to present LBD data for ease of exposition and will often supplement these with plots of the distributions of values to help disambiguate cases where information is lost due to the transformation and to give a better sense of the density of points. Additionally, points in the simplex will be colored according to their position in lbd space, with the red, green, and blue color components corresponding to L, B, and D respectively.

## IV. LBD Distributions

The LBD representation of a graph gives a concise summary 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. This kind of local structure we call the fine structure of a graph. 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. 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 subgraphs centered 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

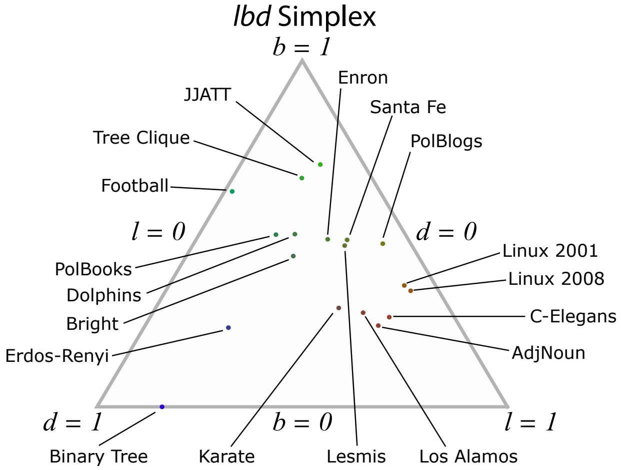


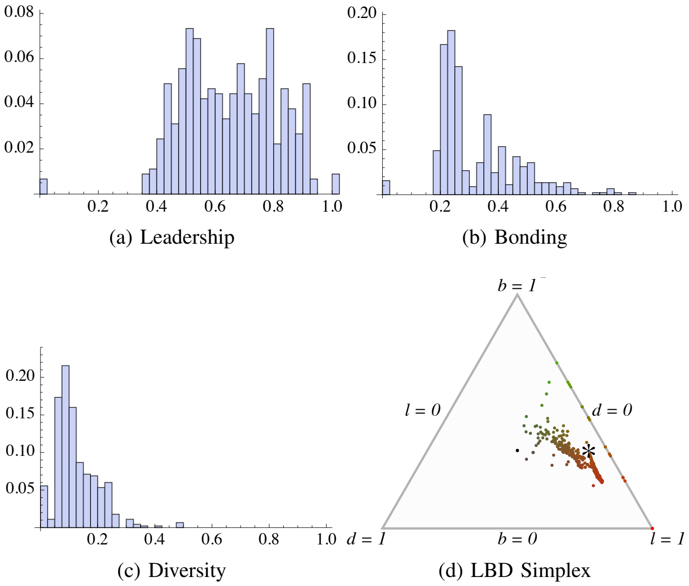
Fig. 1. Graphs of different networks are scattered throughout the LBD space. Here, to clarify, their positions are projected onto the (1,1,1) plane (i.e. the lbd simplex.)

with heterogenous fine structure, whereas a ring is an example of a homogenous graph.

Similar to work on degree distributions and motif analysis which measure the local connectivity and the presence of local subgraph structure across a graph respectively, we represent the fine structure of a graph as a distribution of LBD values. A graphs' LBD distribution is a normalized histogram of the LBD scores of all the induced subgraphs centered on each of its vertices. The distribution has a scale parameter, namely the radius of the subgraphs, 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 subgraph formed by the vertex, its neighbors, and all the edges connecting them. Normalizing 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 each induced subgraph will contain the majority 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 generalizes across classes of motif based on their LBD score.

Figures 2 and 3 show and the radius 2 LBD distributions for two email exchange networks. The first is extracted from the Enron email dataset collected a part of the CALO project [12]. Each vertex is an email address in the data set and an edge links two vertices if the email addresses both sent at least one email to each other. Email 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 [13]. Here each



(a) Leadership

(b) Bonding

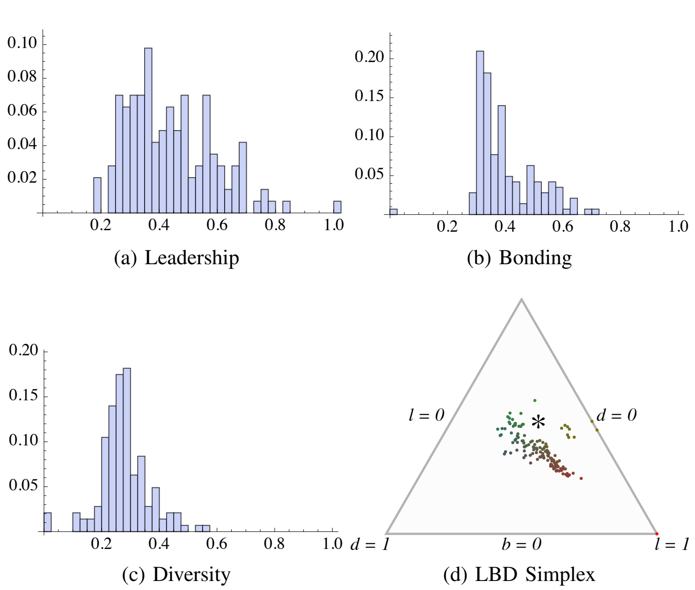
(c) Diversity

(d) LBD Simplex

Fig. 2. LBD distributions and simplex for the Linux-2008 graph at radius 2. 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.)

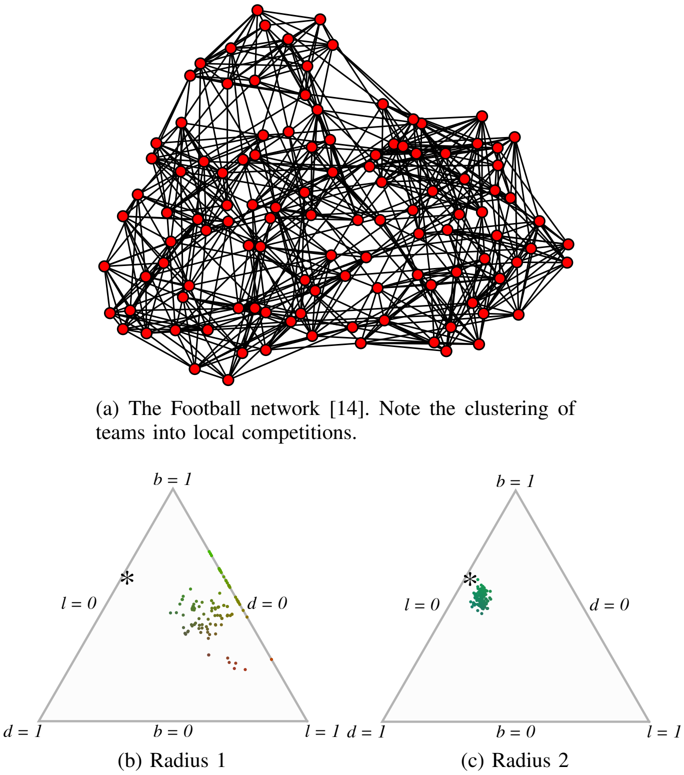
vertex is again an email address, with some email aliases being collapsed into a single vertex. An edge again represents that at least one email was exchanged each way between the two addresses. From the distributions in the two figures we can see that the subgraphs comprising the Linux graph tend to 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 with directly with others in their neighborhood. For Enron the marginally higher bonding suggests more direct communication between people in local neighborhoods and the lower leadership indicates there are fewer people people who are involved in a disproportionately large number of different email 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 don't correspond with each other being joined by a small number of common members. This makes sense for an organization such as Enron where team members might email one another and managers or team leaders serve as communication bridges between teams. It is interesting to note that the full graph LBD score for the Linux graph is close to it's 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.

LBD distributions can look dramatically different across different radii. Figure 4a shows a highly structured graph of football matches between division IA colleges in Fall of 2000 compiled by Girvan and Newman [14], in which each vertex



(a) Leadership
(b) Bonding
(c) Diversity
(d) LBD Simplex

Fig. 3. LBD distributions and simplex for the Enron graph at radius 2. 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.)



(a) The Football network [14]. Note the clustering of teams into local competitions.

(b) Radius 1
(c) Radius 2

Fig. 4. Visualization and lbd simplexes for the Football graph at radii 1 and 2.

is a team and each edge is a match. The general structure is that local teams play one another, forming small bonded subgraphs, and then their winners play one another, linking the subgraphs. Figures 4b and c show the distribution of LBD values at radius 1 and radius 2. At radius 1 we can see the a large proportion of the graph is composed of subgraphs with one or two vertices whose degree is higher than the rest of the vertices in the subgraph. 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 subgraphs, diversity scores tend to be low. This tells us that when we look at just the subgraph 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 neighborhoods include most of the graph, leading to a convergence in LBD scores. Leadership scores become much lower, because now most subgraphs 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.

## V. Comparing Graph Fine Structure

Since the LBD distribution of a graph summarizes 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 subgraphs 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 subgraphs including friends of friends. From our experiments the most interesting results come from analysis at these two radii, particularly radius 2, at which subgraphs become large enough for diversity to be a significant factor.

An issue which was not mentioned in section IV 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 subgraphs. However, for the purposes of ease of comparison we may wish to bin LBD values within discretized regions. The choice of the granularity of this discretization will impact any comparison, since coarser discretizations may place distinct points in the same bin. We chose a compromise between abstraction and fidelity by discretizing LBD space into 0.2 unit length cubes with the result that some graphs may be judged more similar than in the non-discretized case. Our results suggest however that the discretization 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 LBD bin's contribution in the representation by the proportion of subgraphs in the full graph that fall into that bin. 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 subgraph 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 normalize 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 subgraph radius, $r$, and computing histograms, with bin sizes of 0.2, of the LBD scores of the radius $r$ induced subgraphs in each graph. We then normalize 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 [15]–[17] between these two distributions using Euclidean distance as the ground distance. Finally we normalize by the maximum distance in the distcretized 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 [1] 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 modeling a wide variety of phenomena, from social networks and email traffic to football match-ups and neural networks. Table I gives an overview of the graphs included in the analysis, showing the number of vertices, edges, edge probability, and full graph LBD scores. 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 can be seen in figure 5, 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. 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. We will discuss this result further in section VII.

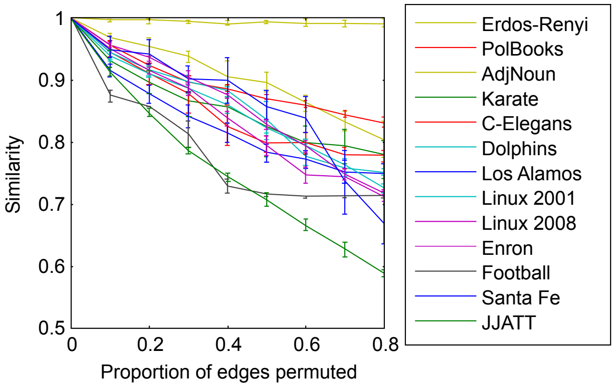


Fig. 5. Radius 2 self-similarity under random edge permutation.

## VI. 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 [27]. 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.

We performed our clustering analysis on the same set of graphs listed in table I. Figure 6a shows the pairwise similarity between each graph in the set computed with a radius of 2. By contrast, 6b 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 IV. 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 7a 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, indicated on the horizontal axis, indicating that two clusters were chosen to be merged at that similarity threshold. The names of graphs derived from

TABLE I

Vertices, edges, edge density, and $LBD$ scores for the analyzed graphs.

| Graph | $\|V\|$ | $\|E\|$ | $P(E)$ | $L$ | $B$ | $D$ | Type | Source |
|---|---|---|---|---|---|---|---|---|
| LosAlamos | 30 | 78 | 0.1793 | 0.6946 | 0.3683 | 0.2923 | Coauthorship | [18] |
| Karate | 34 | 78 | 0.1390 | 0.3996 | 0.2557 | 0.2402 | Social | [19] |
| Dolphins | 62 | 159 | 0.0841 | 0.1164 | 0.3088 | 0.1959 | Social | [20] |
| Enron | 143 | 623 | 0.0614 | 0.2377 | 0.3591 | 0.1455 | Email | [12] |
| Santa Fe | 116 | 174 | 0.0261 | 0.1681 | 0.2200 | 0.0683 | Coauthorship | [14] |
| JJATT | 263 | 998 | 0.0290 | 0.1362 | 0.4905 | 0.0744 | Social | [21] |
| Linux 2001 | 302 | 749 | 0.0165 | 0.2510 | 0.1534 | 0.0333 | Email | [13] |
| Linux 2008 | 450 | 2122 | 0.0210 | 0.3413 | 0.1929 | 0.0388 | Email | [13] |
| Bright | 54 | 175 | 0.1223 | 2.5947 | 0.3770 | 0.2634 | Semantic | [18] |
| Lesmis | 77 | 254 | 0.0868 | 0.3972 | 0.4989 | 0.1755 | Literature | [22] |
| PolBooks | 105 | 441 | 0.0808 | 0.1627 | 0.3484 | 0.1877 | Economic | [23] |
| AdjNoun | 112 | 425 | 0.0684 | 0.3799 | 0.1569 | 0.1320 | Semantic | [24] |
| Football | 115 | 613 | 0.0935 | 0.0120 | 0.4072 | 0.2355 | Sports | [14] |
| C-Elegans | 297 | 2148 | 0.0489 | 0.4066 | 0.1807 | 0.1106 | Neural | [25] |
| PolBlogs | 1490 | 16750 | 0.0151 | 0.2210 | 0.2260 | 0.0327 | Citation | [26] |

social data, such as email 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.

Looking at the clusters formed by the fine structure analysis it is interesting to 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 email 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 email correspondence graph in the data set. 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 data set 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 subgraphs 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 subgraphs 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.

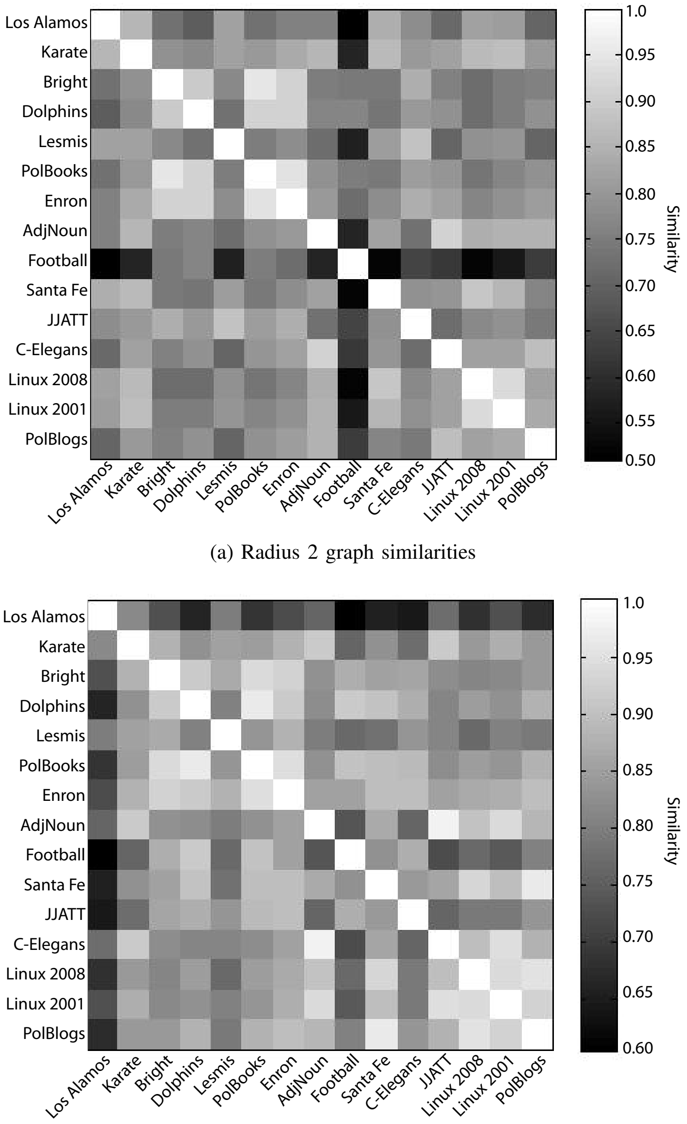
Interestingly, neither measure judges the collaboration networks Santa Fe and Los Alamos to be particularly similar. In the case of fine structure, this is most likely because the small number of vertices in the Los Alamos graph makes its distribution much more sparse along the leadership axis than the 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 email.

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 characterization 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 subgraphs, Dolphins, which is from non-human social data, and the Enron email 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.

## VII. 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? In the case of the Linux email graphs it is reasonable to suggest that their similarity is due to a common generative process that produced them. Further suggestive evidence for fine structure



(a) Radius 2 graph similarities

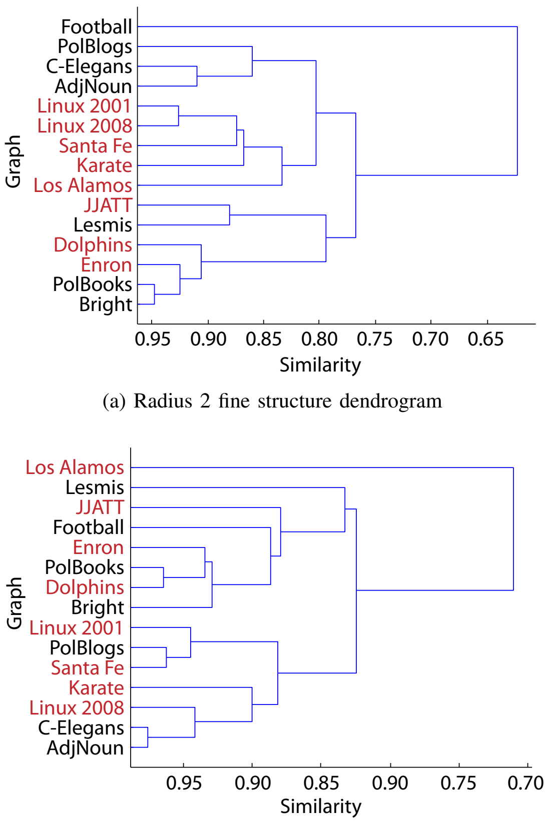
(b) Full graph similarities

Fig. 6. Radius 2 and full graph similarities.

similarity being tied to a graph's generative process comes from the result obtained for Erdős-Rényi random graphs, where our random edge permutation transformations did not significantly impact the similarity of the original graph to its transformation. Note that what is actually being done when we permute edges in this way is transforming one Erdős-Rényi random graph into another instance of an Erdős-Rényi random graph. We have observed from simulation that full graph LBD scores for Erdős-Rényi random graphs tend to lie in a similar region in LBD space and now we have evidence that their LBD distributions likewise tend to be similar. 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 needs to be done to prove this.

Our fine structure analysis gave us evidence that the generative process that produces graphs representing the same phenomena, for instance email 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 organizational 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 email correspondence. As mentioned in the previous section we can draw similar conclusions for collaboration graphs. On the other hand, the fine structure similarity between the two Linux email graphs gives evidence that a generative process that is more organic than that which produced the Football or Erdős-Rényi graphs can indeed give rise to graphs which have similar fine structures and as such should be amenable to empirical study.

Our conclusion is that the specific conditions under which the phenomena that a graph models takes place can be more crucial for its fine structure characteristics than the general class of phenomena that the graph represents. A key challenge for further research then is to characterize 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 model by providing a



(a) Radius 2 fine structure dendrogram

(b) Full graph dendrogram

Fig. 7. Hierarchical clustering dendrograms based on radius 2 and full graph similarity. Graphs with red names are derived from social data.

method for judging the similarity between the fine structure of an empirically observed graph and graphs produced by a proposed model.

## VIII. Conclusion

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 generalizes the idea of computing a distribution of motif subgraphs within a graph by abstracting the structure of subgraphs to leadership, bonding, and diversity scores. These features summarize structural features 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 subgraphs for which the 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 email 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 analyzing that their fine structural similarity did not seem to be dependent upon the phenomena that the graphs were modeling. This led us to conclude that idiosyncratic features of organizations were likely to have more influence on a graph's fine structure than broad commonalities between people's email correspondence or collaborative research behavior. Furthermore our analysis showed that graphs can be judged similar by their full graph structure and yet dissimilar by their fine graph structure, emphasizing the importance of choosing the granularity of analysis at which a similarity judgement is to be made. The results of our analysis on the Linux graphs suggest that common generative processes lead to similar fine structure. This is also borne out by our analysis of the self-similarity of Erdős-Rényi random graphs under permutation. Our technique is a useful tool both for comparing empirical graphs and for comparing the fine structure of graphs produced by a proposed generative model to the empirically observed graphs that they are seeking to explain.

## Acknowledgments

The authors would like to thank Omprakash Gnawali for providing data for the Linux kernel email correspondence networks and the AFOSR for their support in this research.

## References

[1] M. Peabody, "Finding groups of graphs in databases," Master's thesis, Drexel University, Philadelphia, 2002.

[2] D. McWherter, "Approximate variations of graph matching and applications," Master's thesis, Drexel University, Philadelphia, 2001.

[3] C. J. Anderson, S. Wasserman, and B. Crouch, "A p* primer: Logit models for social networks," *Social Networks*, vol. 21, pp. 37–66, 1999.

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

[5] K. M. Borgwardt, "Graph kernels," Ph.D. dissertation, Ludwig Maximilians University, Munich, 2007.

[6] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, pp. 824–827, 2002.

[7] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.

[8] R. Read and R. Wilson, *An Atlas of Graphs*. Oxford Press, 1998.

[9] W. Richards and N. Wormald, "Representing small group evolution," in *Proceedings of the IEEE Conference on Social Computing*, 2009, p. 232.

[10] L. C. Freeman, "Centrality in social networks: conceptual clarification," *Social Networks*, vol. 1, pp. 215–239, 1978.

[11] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press, 1994.

[12] W. W. Cohen. (2009) Enron email dataset. [Online]. Available: http://www.cs.cmu.edu/ enron/

[13] O. D. Gnawali, "Linux kernel email communication networks from january 2001 and 2008," Personal Communication, 2009.

[14] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[15] O. Pele and M. Werman, "A linear time histogram metric for improved sift matching," in *Proceedings of the European Conference on Computer Vision*, 2008.

[16] ——, "Fast and robust earth mover's distances," in *Proceedings of the International Conference on Computer Vision*, 2009.

[17] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Proceedings of the International Conference on Computer Vision*, 1998.

[18] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, pp. 814–818, 2005.

[19] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, pp. 452–473, 1977.

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

[21] S. Atran, S. Bennett, A. Fatica, J. Magouirk, D. Noricks, M. Sageman, and D. Wright. (2008) John Jay & ARTIS Transnational Terrorism (JJATT) dataset. [Online]. Available: http://doitapps.jjay.cuny.edu/jjatt/

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

[23] V. Krebs. (2003) Books about US politics dataset (unpublished). [Online]. Available: http://www.orgnet.com/

[24] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physics Review E*, vol. 74, 2006.

[25] J. G. White, E. Southgate, J. N. Thompson, and S. Brenner, "The structure of the nervous system of the nematode c. elegans," *Philosophical Transactions of the Royal Society*, vol. 314, pp. 1–340, 1986.

[26] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 US election," in *Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem*, 2005.

[27] M. H. Dunham, *Data Mining: Introductory and Advanced Topics*. New Jersey: Prentice Hall, 2002.