Investigating the Fine Grained Structure of Networks

by

Owen Macindoe

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2010

Author .....
Department of Electrical Engineering and Computer Science
May 21, 2010

Certified by .....
Whitman Richards
Professor of Cognitive Science
Thesis Supervisor

Accepted by .....
Terry P. Orlando
Chair, Department Committee on Graduate Students

Investigating the Fine Grained Structure of Networks

by

Owen Macindoe

Submitted to the Department of Electrical Engineering and Computer Science
on May 21, 2010, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

In this thesis I explore a novel representation for characterizing a graph's fine grained structure. The key idea is that this structure can be represented as a distribution of the structural features of subgraphs. I introduce a set of such structural features and use them to compute representations for a variety of graphs, demonstrating their use in qualitatively describing fine structure. I then demonstrate the utility of this representation with quantitative techniques for computing graph similarity and graph clustering. I show that similarity judged using this representation is significantly different from judgements using full graph structural measures. I find that graphs from the same class of networks, such as email correspondence graphs, can differ significantly in their fine structure across the institutions whose relations they model, but also find examples of graphs from the same institutions across different time periods that share a similar fine structure.

Thesis Supervisor: Whitman Richards
Title: Professor of Cognitive Science

Acknowledgments

To Miss Sophie Ella Mackey, by far the dishiest vertex in my ego-centric subgraph for $r = [1, \infty]$.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem description

This thesis presents a novel method for representing the structure of graphs based on the structural features of their constituent subgraphs and demonstrates its utility as a tool for network analysis.

From social networks in Facebook to connectivity maps of neurons, scientists in a broad range of disciplines are faced with the challenge of analyzing the structure of networks. Such an analysis is crucial in explaining processes occurring in the systems they model, such as the spread of STIs in a network of sexual partners or the flow of information through a network of social acquaintances. Network structure is typically modeled using graphs, where vertices model the objects of interest and edges model the relations between those objects, and in this thesis I will frequently refer to the structure of a network when I mean the structure of the graph that models it. Even with state of the art tools for visualizing graphs, making judgements about network structure is challenging, particularly for graphs having more than a hundred vertices. For example, fundamental questions, such whether two graphs are structurally identical, are hard to decide visually and are suspected to be computationally intractable. In this thesis I will be particularly concerned with the question of what the structural patterns of the local neighborhoods centered around any given vertex in a graph are and how these patterns vary across the graph's vertices. This local

structure I call the "fine grained structure" or simply "fine structure" of a network. The key problem this thesis addresses is how this structure can be characterized, and once characterized, how can it be compared across graphs.

In this thesis I present a useful and compact representation for the structure of graphs. By useful I mean that it quantifies features of a graph that are of interest for network analysis and which reveals the homogeneity or heterogeneity of fine structure across a graph. This allows for tasks such as computing structural similarity or graph clustering. By compact I mean that the representation renders the fine structure of a graph more amenable to visualization and analysis by humans than visualizations produced by standard graph layout algorithms, scaling well to graphs with hundreds or thousands of edges.

1.2 Approach

My approach to analyzing a graph proceeds by introducing a set of features, namely *leadership*, *bonding*, and *diversity*, which summarize structural properties of a network that are motivated by concerns in social domains [32]. The graph's structure is represented by the distribution of these features computed for subgraphs centered on each vertex in the graph. The granularity of the representation can be controlled by varying the radius of these subgraphs, and I show that choice of granularity has a strong impact on the structural properties revealed. I also demonstrate that performing a multi-scale analysis by comparing distributions across radii can be informative.

Visualizing the distribution of subgraph features reveals the homogeneity and heterogeneity of a graph's fine structure and allows the characterization of the graph by the proportion of subgraphs that are more or less dominated by the above three features. These visualizations reveal elements of local structure that are hard to detect by inspection of an image of a graph's vertices and edges. I show and discuss distribution visualizations for graphs from a wide variety of domains.

With graphs represented as distributions of the structural features of their subgraphs it is possible to compute the earth mover's distance between distributions for

two graphs, giving a measure of their structural similarity [33]. I demonstrate that this technique for judging similarity yields intuitive results by showing that graph self-similarity decreases with increasing edge permutation. Similarity computed in this way can be used to cluster graphs based on their fine grained structure using standard clustering algorithms, a technique which I demonstrate and which reveals both similarities and dissimilarities between graphs from the same domain, as well as interesting some cross domain similarities.

1.3 Outline of thesis

This thesis proceeds with the following structure:

Chapter 2 introduces the formalism for describing graphs that will be used throughout this thesis, gives an overview of structural features of graphs common in the network science literature, and reviews previous approaches to graph structure comparison.

Chapter 3 reviews the three key structural features, *leadership* (L), *bonding* (B), and *diversity* (D), used in the fine structure representation. For a variety of networks I give examples of the joint L, B, and D (LBD) scores that describe them, and present them via a simplex visualization that I will use throughout the thesis.

Chapter 4 introduces LBD distributions, which represent the fine grained structure of graphs, and are this thesis' key contribution. I discuss the computation of these distributions and give examples of LBD distributions for three sets of graphs: manually constructed graphs designed for didactic purposes, graphs derived from human and non-human social networks, and graphs derived from networks of other kinds.

Chapter 5 demonstrates using the earth mover's distance between the LBD distributions of two graphs to judge their structural similarity. I discuss results obtained for computing the similarity between the graphs introduced in chapter 4 and use this similarity measure to cluster graphs based upon the similarity of their fine grained structure.

Chapter 6 summarizes my key findings, namely that a focus on fine grained structure does indeed reveal important differences between graphs that a full graph approach does not, that the technique introduced in chapter 5 is a reasonable measure of structural similarity, and that evidence for fine structure similarity between graphs modeling a common process across institutions or domains is weak, yet there is evidence of such similarity in graphs generated across different time periods within the same institution. I then suggest future work in developing graph representations based on the distribution of the structural features of their subgraphs, extensions of leadership, bonding, and diversity, and new directions suggested by the similarity and clustering results.

Chapter 2

Background

2.1 Graphs

Systems taking the form of networks are typically modeled by mathematical objects called *graphs*, composed of *vertices*, which model objects in the network, and *edges*, which model the relations between these objects [17]. In this section I will introduce some basic concepts from graph theory that will be useful for discussing graph structure.

Formally, a graph G is a pair of sets, $G = (V, E)$, where $V = \{v_1, ..., v_n\}$ are the n vertices of the graph and $E = \{e_1, ...e_m\}$ are the m edges of the graph. Each edge $e \in E$ is itself a pair $(u, v) \in V \times V$, indicating that the relation modeled by the edge holds between u and v. If $(u, v) \in E$ then the vertices u and v are said to be *adjacent* or, equivalently, they are *neighbors*. The *degree* of a vertex in a graph is the number of vertices to which it is adjacent. Edges in a graph can be *directed*, that is $e_1 = (u, v)$ is distinct from an edge $e_2 = (v, u)$, however in this thesis I will deal only with *undirected graphs*, in which such edges are equivalent to each other.

The *order* and *size* of a graph are, respectively, the number of vertices, $|V|$, and edges, $|E|$, in the graph. A *walk* in a graph is a sequence of vertices and edges such that each edge joins the vertices that immediately precede and follow it. A *path* in a graph is a walk in which no vertex is repeated. The length of a walk or a path is the number of edges ocurring in it . Walks and paths are often referred to as *k-walks*

or *k-paths* when they contain k edges. A *cycle* is a path with only one repetition, namely the first and last vertices are the same. A *k-cycle* is a cycle with k edges. A 3-*cycle* is also called a *triangle*.

The *distance* between two vertices in a graph is the length of the shortest path between them and such a path is called a *geodesic*. The *diameter* of a graph is the length of the longest geodesic between any two vertices in a graph. The *radius* of a graph is the shortest distance such that there is some vertex for which all other vertices in the graph are within that distance of it. The mean geodesic distance between any two vertices in a graph is called the *characteristic path length* (CPL) of the graph.

A *component* of a graph is a maximal subset of its vertices such that a walk exists between each pair of vertices. If a graph contains only one component it is said to be *connected*. A graph $G' = (V', E')$ is said to be a *subgraph* of $G = (V, E)$ if $V' \subseteq V$ and $E' \subseteq E$. Such a subgraph is additionally an *induced subgraph* if for every pair of vertices $u, v \in V'$, $(u, v) \in E'$ if and only if $(u, v) \in E$. Two graphs G and G' are *isomorphic* if a bijective function $f : V \to V'$ exists such that $(u, v) \in E \iff (f(u), f(v)) \in E'$ for all $u, v \in V$. Intuitively, two graphs are isomorphic if a vertex mapping scheme can be found that preserves adjacencies between vertices.

Several special classes of graphs will be of interest in this thesis, namely dipoles, trees, complete graphs, stars, regular graphs, bipartite graphs, and Erdős-Rényi random graphs [17, 31, 10]. *Dipoles* are graphs consisting solely of two vertices joined by one edge. *Trees* are connected graphs that contain no cycles. *Complete graphs* are graphs in which there is an edge between every pair of vertices. A *star* contains one central vertex that is adjacent to all other vertices and no other vertices are adjacent to each other. A *k-star* is a star with k vertices. A *regular* graph is one in which all vertices have the same degree. A *k-partite* graph is one whose vertices can be separated into k disjoint sets such that no two vertices in the same set are adjacent. 2-partite graphs are also called *bipartite* graphs. *Erdős-Rényi* or *Poisson random graphs* are graphs constructed by starting with a vertex set V, selecting a coin weight

p in the range $[0, 1]$, flipping the coin once for each possible edge position, and adding the edge if the coin comes up heads.

The two most common data structures used to represent graphs are *adjacency matrices* and *edge lists*. An adjacency matrix for an undirected graph is a $|V| \times |V|$ matrix in which an entry (i, j) is 1 if $i \neq j$ and v_i is adjacent to v_j, otherwise the entry is 0. Specifying $i \neq j$ forbids a vertex from having a self-connecting edge. The matrix representation makes computations such as finding all paths of a given length easy, but can be wasteful of memory for graphs with few edges. The edge list representation simply stores V and E as lists. This representation trades computational efficiency in some algorithms for reduced memory requirements.

2.2 Structural features of graphs

Over the past decade there has been a convergence of interest in graph structure across fields, resulting in a rich literature on structural features of graphs across many domains. In this section I will review a selection of these features.

Edge probability or *network density*, $pE(G)$ is a commonly used macro feature of graphs [24]. It is the proportion of possible edges in a graph that actually appear in it and is given by the equation:

$$pE(G) = \frac{2 \times |E|}{|V|(|V| - 1)} \tag{2.1}$$

A common feature of social networks studied by sociologists and social psychologists is their tendency towards *triadic closure* or *transitivity* [38]. In an acquaintance network, this is the degree to which people who share a mutual acquaintance are themselves acquaintances. A measure of this, called the *fraction of transitive triples*, and also one of several features called the *clustering coefficient* [24], is widely used in the sociological literature and for undirected graphs is is given by the equation:

$$C(G) = \frac{6 \times \text{triangles in graph}}{\text{2-paths in graph}} \tag{2.2}$$

A clustering coefficient variant introduced by [39] instead defines triadic closure locally, computing a clustering coefficient for each vertex:

$$C_i(G) = \frac{\text{triangles connected to } v_i}{\text{3-stars centered on } v_i} \tag{2.3}$$

With $C_i = 0$ in the case that the denominator is zero. Recall that a 3-star is two non-adjacent vertices connected to a third vertex, which is the center of the star. The mean clustering coefficient across all vertices can then be taken as an indicator of triadic closure across the whole graph. The distinction between local and global graph structure will feature prominently in the graph representation presented in this thesis.

The notion that some vertices and edges are more important than others has led to various centrality features being proposed throughout the literature. One example is *betweenness centrality*, which is the number of geodesics between a graph's vertices that run through a given edge or vertex [23]. The distribution of betweenness centrality in a graph has implications for its resilience, with graphs with high-skewed distributions being vulnerable to disruption through the removal of key vertices or edges. If edges represent flows of resources, the high betweenness centrality of an edge can also indicate a bottleneck [36].

Over the last decade the *degree distribution* of graphs has been of great interest to researchers [4, 3, 8]. A graph's degree distribution is a histogram of the degrees of each vertex in the graph. A key finding across domains as diverse as citation networks and metabolic networks has been that the degree distribution of many networks follows a *power law*, with the number of vertices in a graph having degree k being proportional to $\frac{1}{k^c}$ for some constant c [3]. This contrasts with the binomial distribution associated with Erdős-Rényi random graphs, in that the number of high degree nodes is substantially larger than would be the case in a random graph [10]. Networks with power law degree distributions are called *scale free* networks and models of their formation have been well developed [3, 8]. The key contribution of research in this area has been to show that preferential attachment processes, where new vertices being added to

a graph form edges preferentially with vertices with a high degree, thus causing the high degree vertex's degree to become yet higher, robustly predict power law degree distributions under a wide range of variations. The ubiquity of this feature, combined with the intuitive appeal of the models explaining it, make it an outstanding example of success in network structure analysis.

A variety of methods for detecting communities in social networks using structural features have been proposed. These include using hierarchical clustering based on vertex-vertex connection strength [11], detecting the borders of communities via edges with high betweenness centrality [15], spectral techniques [26], and comparing the density of edges within proposed communities to that of the graph as a whole [25]. Despite the interest in this topic, a lack of ground truth data for communities in social networks has made evaluating these techniques challenging.

Motif analysis [22, 37], computes the frequency of the occurrence of small induced 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 gives a characterization of a graph's structure in terms of its statistically significant subgraphs. A key question for this kind of analysis is what particular graphs these motifs should be. This question will be influential in the development of the graph representation presented in this thesis.

2.3 Graph similarity

Relevant to understanding a 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, judging graph similarity is key.

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 [6]. Unfortunately the problems with edit-distance are twofold. First there are many possible kinds of edit operations, including 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 problem of detecting graph isomorphism, which has no known polynomial time solution. It's clear that we will have to accept some level of approximation in any similarity measure for the sake of tractability.

Graphs are often informally compared using the full-graph structural features discussed in section 2.2 and by distributions of local features such as vertex degree, vertex centrality, or motif appearances. These kinds of comparisons are often presented in tabular form or as visualizations of distributions, usually with the intent to illustrate qualitative differences rather than to compute a quantitative similarity value.

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 [29, 21]. This was demonstrated in [29] by cloning graphs, randomly permuting their copies, and showing that their spectral distance increases as a function of the amount of permutation. Spectral techniques have two weaknesses however, the first being the existence of relatively rare isospectral graphs, which share eigenvalues despite having quite different structure and therefore can erroneously be judged similar. More relevant for this thesis 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 socially motivated features.

Graph kernels have become a popular approach to comparing graph structure. They are a broad class of functions that map pairs of graphs, (G, H), to a typically high dimensional inner product space [5]. This form of manipulation known as

the *kernel trick* is used extensively in machine learning techniques such as Support Vector Machines [34]. Such techniques can, in turn, be used to construct classifiers which separate graphs into classes which are in some sense self-similar. Graph kernel techniques range widely, but two illustrative examples are exhaustive pairwise isomorphism testing on all subgraphs of G and H, which is NP-hard to compute [14], and more efficient methods based on subtree comparison [35].

Chapter 3

Summarizing graph structure

In section 2.2 I reviewed a range of features that have been used to measure aspects of graph structure. Because this thesis will be largely concerned with the structure of graphs derived from social network data, I have selected three features to use in the graph representation presented here that are socially motivated. These features, first introduced in [32], are characterized as *leadership* (L), *bonding* (B), and *diversity* (D) . This thesis will use LBD triples to represent undirected graphs as points in the LBD feature space. In this section I review these features and present examples of L, B, and D values computed for various graphs.

3.1 Leadership

Leadership is a measure of the extent to which the edge connectivity of a graph is dominated by a single vertex [12]. It is given by equation 3.1, in which $|V|$ is the order of the network, d_i is the degree of vertex i and d_{max} is the maximum degree of any vertex in the graph. It is the mean difference between the degree of the highest degree vertex and that of each other vertex in the graph. Leadership is maximal (i.e 1) in a star graph (one vertex of degree $|V| - 1$ with all other vertices of degree 1) and zero for regular graphs, in which all vertices have the same degree (e.g. a complete graph or a ring). In equation 3.1 the denominator is the maximal sum, which normalizes L to $[0, 1]$. In an acquaintance 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.

$$L = \frac{\sum_{i=1}^{|V|}(d_{max} - d_i)}{(|V| - 2)(|V| - 1)} \tag{3.1}$$

3.2 Bonding

Bonding has already been presented in section 2.2 under the name of clustering coefficient and is given by equation 3.2. Because this term is ambiguous in the literature I have followed [32] and adopted the name bonding for the sake of clarity. 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. 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 an acquaintance network a high bonding means that if two people are share a mutual acquaintance, then it is likely that they themselves are acquainted.

$$B = \frac{6 \times \text{triangles in graph}}{\text{2-paths in graph}} \tag{3.2}$$

3.3 Diversity

Diversity, given by equation 3.3, is a measure based on the number of pairs of dipoles in a graph whose end vertices are not adjacent, and hence are disjoint. A normalization is imposed by the maximal count, which occurs for the complement of the complete bipartite graph and the square root of the ratio scales the measure into a range similar to L and B [32]. $D = 0$ for $|V| < 4$ and possible values lie in the range $[0, 1]$. Diversity is high in graphs that are not densely connected, such as rings, but also in graphs where separate graph regions are joined by a relatively small number of bridging edges. In an acquaintance network a high diversity indicates that separate communities exist, where people from one community have no direct ties with people

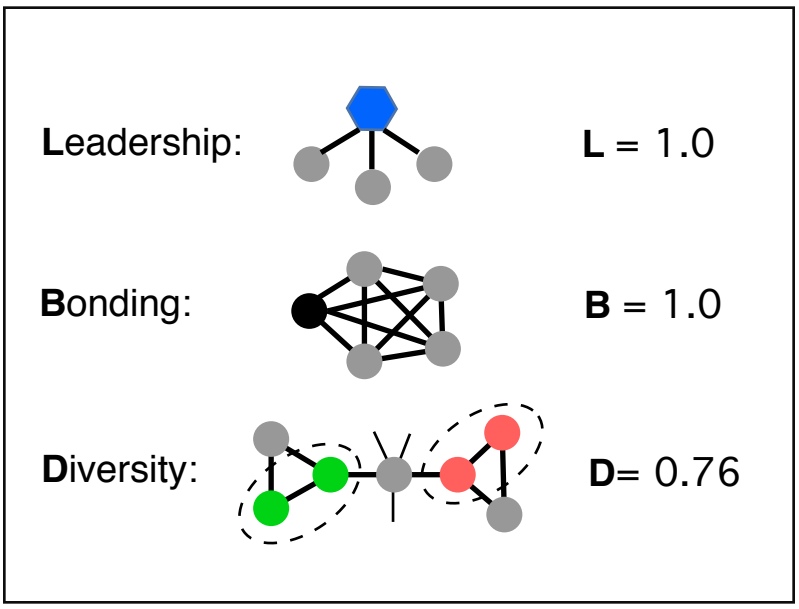


Figure 3-1: Canonical examples of graphs with high leadership, bonding, and diversity scores. Highlighted vertices are a "leader", a member of a clique, and the end points of two disjoint dipoles.

in another, whereas a low diversity indicates that people are generally all connected to one another.

$$D = \sqrt{\frac{\text{disjoint dipoles in graph}}{(\frac{|V|}{4}(\frac{|V|}{2}-1))^2}} \tag{3.3}$$

Figure 3-1 shows canonical examples of graphs with high L, B, and D scores, namely a star, a complete graph, and a "bow-tie". Further discussion of L, B, and D can be found in [32].

3.4 Analyzing graphs using leadership, bonding, and diversity

I analyze 18 graphs in this thesis, broken down into three classes. The first are graphs constructed for didactic purposes to demonstrate using LBD distributions to characterize graph structure for graphs whose structure is well understood. The second are graphs constructed from social data, including friendship networks, scientific collab-

orations, and email correspondence. The third are graphs formed from other kinds of relational data, from co-occurances of words in text to neural connectivity maps. Table 3.1 summarizes the semantic content behind the non-constructed graphs, giving the objects and relations that vertices and edges in the graphs represent. Additionally, table 3.2 gives some key summarizing features for each of the graphs, including their LBD scores. It is important to note that where the original edge data for these networks was directed I have abstracted it to be undirected, where edges may have originally been weighted this too has been abstracted away, and likewise where vertices had class labels. Thus there is necessarily a loss of structural information in the analysis presented in this thesis which must be considered in any of the results that follow. Edge direction has been abstracted from the C-Elegans and PolBlogs graphs. Edge weights have been abstracted from the C-Elegans and Lesmis graphs. Vertex labels have been abstracted from the AdjNoun, PolBlogs, and PolBooks graphs.

Table 3.1: A summary of the semantics for the edges and vertices of each graph.

Graph	Vertices	Edges	Source
LosAlamos	Scientists	Collaboration on a paper	[28]
Karate	Club members	Friendship	[41]
Dolphins	Dolphins	Significant time in proximity	[20]
Enron	Enron employees	One email exchanged each way	[7]
Santa Fe	Scientists	Collaboration on a paper	[15]
JJATT	Terrorists	Known association from court records	[2]
Linux 2001 (Jan)	Kernel mailing list members	Email exchanged either way	[16]
Linux 2008 (Jan)	Kernel mailing list members	Email exchanged either way	[16]
Bright	Words	Reported in free association	[28]
Lesmis	Characters in Les Miserables	Co-appearance in scene	[18]
PolBooks	Books	Purchased together	[19]
AdjNoun	Adjectives and nouns	Co-occurence in David Copperfield	[26]
Football	College football teams	Match played	[15]
C-Elegans	Neurons	Neural connectivity	[40]
PolBlogs	Political blogs	Hyperlink between blogs	[1]

Together, L, B, and D summarize a graph's structure along three socially motivated dimensions. Plotting graphs in this space is a first step in determining which graphs are similar to one another. In this thesis I follow the convention introduced

Table 3.2: Summarizing features for the analyzed graphs. The asterisk indicates that the PolBlogs graph is not connected and the reported values are for its largest component.

Graph	$\|V\|$	$\|E\|$	pE	Diameter	CPL	L	B	D
Binary Tree	127	126	0.0157	12	8.3510	0.0082	0.0000	0.0435
Tree Clique	62	496	0.2623	10	5.2512	0.2541	0.9945	0.2570
Erdős-Rényi	115	598	0.0912	4	2.2632	0.0768	0.0853	0.2110
Los Alamos	30	78	0.1793	4	2.0598	0.6946	0.3683	0.2923
Karate	34	78	0.1390	5	2.4082	0.3996	0.2557	0.2402
Dolphins	62	159	0.0841	8	3.3570	0.1164	0.3088	0.1959
Enron	143	623	0.0614	8	2.9670	0.2377	0.3591	0.1455
Santa Fe	116	174	0.0261	15	6.6576	0.1681	0.2200	0.0683
JJATT	263	998	0.0290	13	5.8750	0.1362	0.4905	0.0744
Linux 2001	302	749	0.0165	7	3.1614	0.2510	0.1534	0.0333
Linux 2008	447	2122	0.0213	6	2.7919	0.3435	0.1929	0.0393
Bright	54	175	0.1223	5	2.5947	0.2257	0.3770	0.2634
Lesmis	77	254	0.0868	5	2.6411	0.3972	0.4989	0.1755
PolBooks	105	441	0.0808	7	3.0788	0.1627	0.3484	0.1877
Adj-Noun	112	425	0.0684	5	2.5356	0.3799	0.1569	0.1320
Football	115	613	0.0935	4	2.5082	0.0120	0.4072	0.2355
C-Elegans	297	2148	0.0489	5	2.4553	0.4066	0.1807	0.1106
PolBlogs	1490	16715	0.0151	8*	2.7375*	0.2210	0.2260	0.0327

in [32] of transforming LBD scores into points on the 2D (1,1,1) plane which I will call the *lbd simplex*. 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, which are then plotted in the simplex. This gives a qualitative sense for which features are proportionally dominant for a given graph. For instance graphs close to the $b = 1$ and $d = 1$ corners of the simplex have high L and B relative to D. Figure 3-2 shows the l, b, and d scores for each of the graphs analyzed in this thesis. Throughout this paper I 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 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.

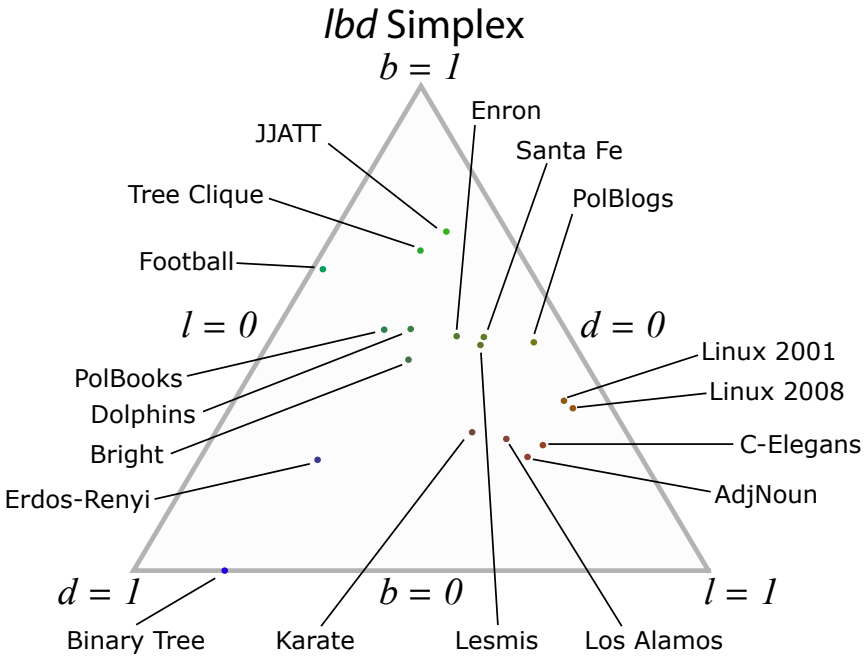


Figure 3-2: *lbd* scores for each graph plotted in the simplex. A graph's position in the simplex indicates the relative magnitude of its L, B, and D scores.

A key thing to note in figure 3-2 is that some graphs, namely Binary Tree and Football, are dominated by one particular score. The reason for this will become clear in the next chapter, but has to do with their structural homogeneity. D scores tend to be low in general for non-constructed graphs, which is reflected in the fact that every such graph appears in the $d < 0.5$ region of the simplex. This is a result of D being sensitive to the ratio of the size of a graph to its order and the fact that most of the non-constructed graphs are quite sparse. A final key point to note is that there is no distinctive region in which only graphs from social data appear. One may naively expect that some distinctive property of social phenomena might lead to social networks having common macro-level structure, but measured in terms of LBD scores this is not the case. This lack of structural distinctiveness for social networks will be a recurring theme throughout the rest of this thesis.

Chapter 4

LBD distributions

4.1 Motivation

The LBD representation of a graph gives a summary of properties of the graph as a whole. But consider the case of a graph with multiple structurally distinct regions, an extreme example of which might be a series of complete subgraphs joined together in a chain by bridging edges. Ideally a representation would 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 region in LBD space. More generally, one 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 with heterogenous fine structure, whereas a ring is a canonical 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, the fine structure of a graph can be represented 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, r, of the subgraphs, which controls the coarseness of the analysis. For example, to compute the $r = 1$ 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, where the normalized value of each bin gives the proportion of the graph's subgraphs that have LBD scores in that bin's region. 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 eventually contain all of the graph's vertices and edges.

An 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 will map to the same value; for instance all star graphs, regardless of size, map to $L = 1$, $B = 0$, $D = 0$. An LBD distribution then is akin to a motif distribution which generalizes across classes of motif based on their LBD scores.

4.2 Method

In this chapter I will present LBD distributions computed for each graph in the sets introduced in section 3.4. For each graph I have computed the distribution of LBD scores for the induced subgraphs centered on each of the graph's vertices and present these data as a normalized histogram for each feature dimension L, B, and D, along with *lbd* simplex plots containing a point for each vertex's subgraph. To demonstrate the effect of the scale parameter, r, on the representation I present and discuss the results at radius 1, 2, and 3.

The first set of graphs, the constructed graphs, is intended to show clearly how fine structure is revealed through this analysis by focusing on graphs that have well understood structures. For the second and thirds sets, the social and other graphs respectively, the structure is less clear cut and the analysis is necessarily more conservative and qualitative. Throughout this section I will be particularly concerned with the $r = 2$ distributions, which, as I will discuss later, offer a good compromise between fine granularity and representativeness. For each graph I include a visual-

ization of its vertices and edges using the GEM algorithm from [13], along with the $r = 2$ *lbd* simplex, and I highlight the key features of the graph's LBD distribution across radii. In each simplex visualization the full graph's *lbd* location is represented with an asterisk as a point of reference. For reasons of space, the LBD distributions and *lbd* simplex visualizations at radii other than 2 are only included with the main text for the constructed graphs. The remainder can be found in appendix A.

4.3 Constructed graphs

4.3.1 Binary Tree

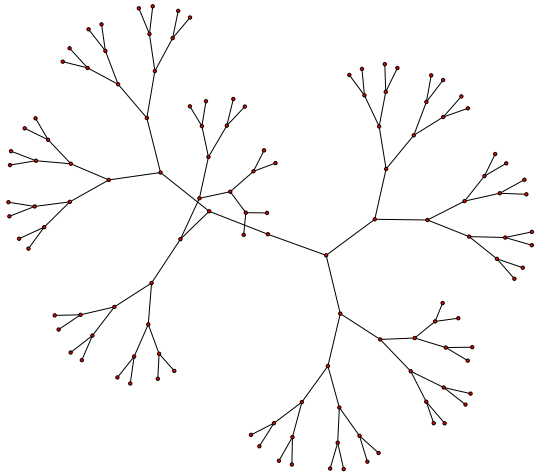
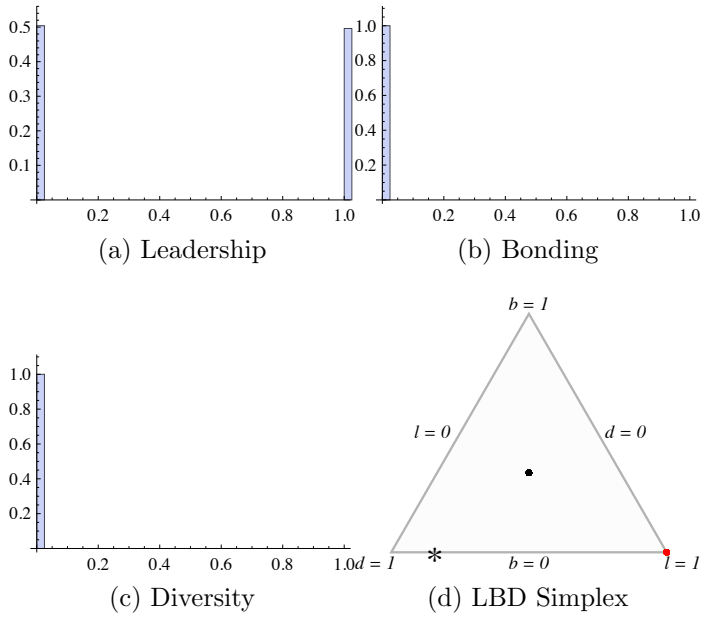


Figure 4-1: Binary Tree graph.

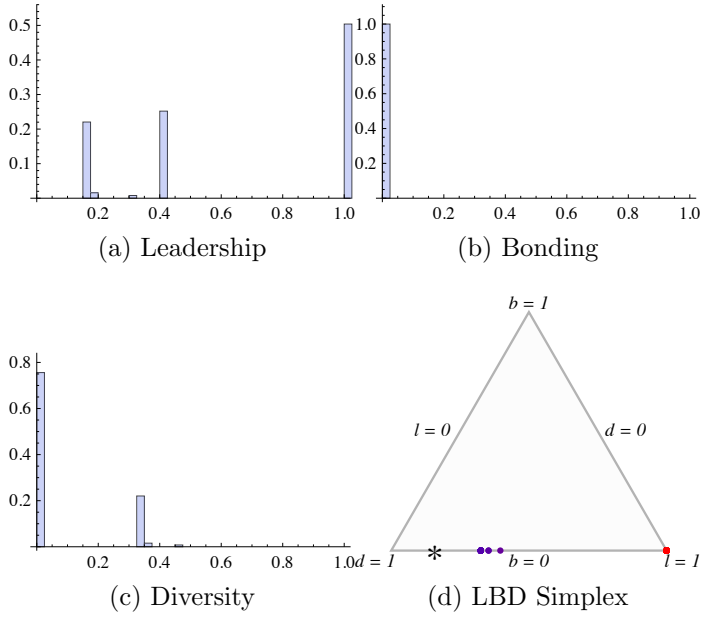
The Binary Tree graph is a 127 vertex binary tree with a depth of 6. The full graph's low absolute D score reflects this feature's sensitivity to the ratio of a graph's size to its order. Viewed in relative terms though, the D score dominates its B and L scores, with no triangles and the almost half the vertices sharing the maximum degree of 3.



(a) Leadership (b) Bonding

(c) Diversity (d) LBD Simplex

Figure 4-2: LBD distributions and *lbd* simplex for the Binary Tree graph at radius 1.

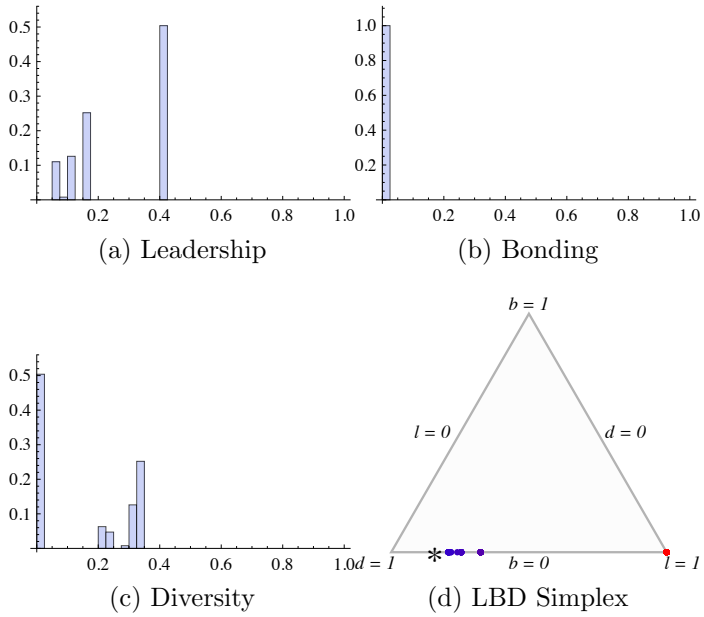


(a) Leadership (b) Bonding

(c) Diversity (d) LBD Simplex

Figure 4-3: LBD distributions and *lbd* simplex for the Binary Tree graph at radius 2.

At $r = 1$ there are only two kinds of subgraph, the dipoles centered on leaves in the tree, and the 3-stars centered on all other vertices except the tree's root, which is the center of a 2-star. This is clearly visible in the simplex, where the point in the center represents the dipoles and the point at the *lbd*-tuple $(1, 0, 0)$ represents the stars. The distinction is also clearly seen in the distribution of L, with the 0 and 1 bins each containing half of the vertices. B and D scores are all 0 as expected at this radius.



(a) Leadership (b) Bonding

(c) Diversity (d) LBD Simplex

Figure 4-4: LBD distributions and *lbd* simplex for the Binary Tree graph at radius 3.

All of the subgraphs centered on leaves become 3-stars at $r = 2$, and the previous stars now become trees themselves. These trees will differ across the graph, depending on the depth of their central vertex, which controls the relative number of degree 1, 2, and 3 vertices in the local subgraphs, which in turn controls their L scores, as well as the number of disjoint dipoles, which controls their D scores. This can be seen in the L histogram where the score of around 0.3 for the graph centered on the full tree's root can be seen, as well as the scores of around 0.2 for the subgraphs around its immediate children. The remaining two bins represent the subtrees rooted higher and lower in the tree. Of the 96 subgraphs with $D = 0$, 64 are the stars rooted at the tree's leaves, and the remainder are the subgraphs rooted at the parents of the

leaves, whose branches are not deep enough to form a 4-path.

As the local neighborhood subgraphs grow in size we see their L scores decrease as the proportion of vertices sharing the maximum of degree 3 grows. At $r = 3$ the subgraphs centered on the leaves have changed from stars to trees, hence no $L = 1$ subgraphs remain. Conversely we see D increase as the number of disjoint dipoles grows with the size of the subtree graphs, however this trend is fought against by the normalization of D which also grows with the square of the graph size, leading to the low score for the full graph mentioned previously. $r = 3$ is the maximum radius at which $D = 0$ local subgraphs exist, with those subgraphs being the ones centered on the leaves.

4.3.2 Tree Clique

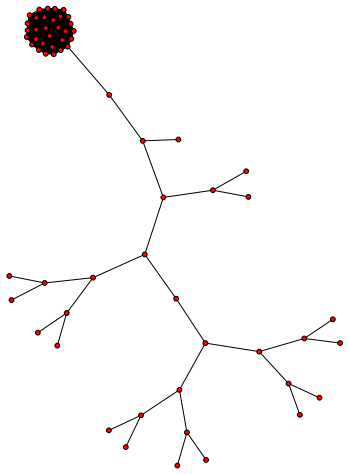
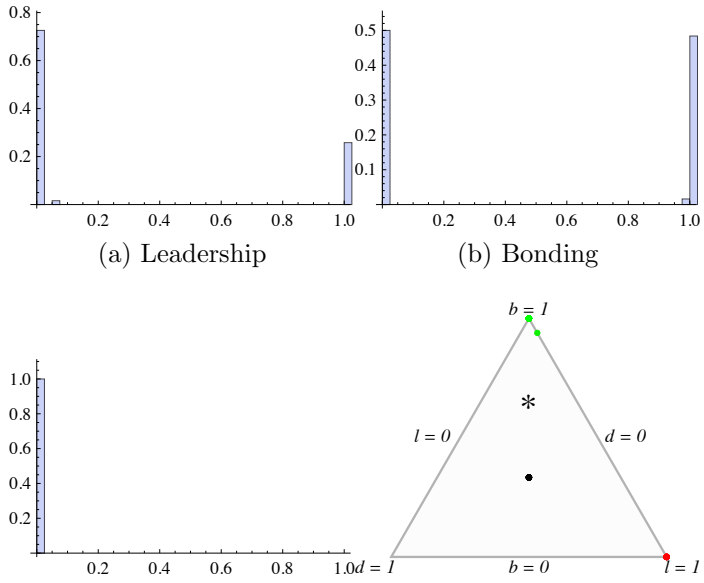


Figure 4-5: Tree Clique graph.

The Tree Clique graph is a 62 vertex graph, combining a binary tree, as in the previous example, with a complete graph. The binary tree section is a 31 vertex binary tree with a depth of 4. Joined to one of the tree's leaves by a single edge is a 31 vertex complete graph. Due to influence of the complete graph, the combined graph has an edge probability of 0.2541 which is the highest among the graphs analyzed in this thesis. Its full graph B score is dominant due the the presence of the complete graph. Its D score is also relatively high due to the graph's large size and the fact that most

dipoles in the complete graph are disjoint from those in the tree. Finally, its L score reflects the fact that vertices in the complete graph have very high degree relative to the vertices in the tree, but are on a par with each other.

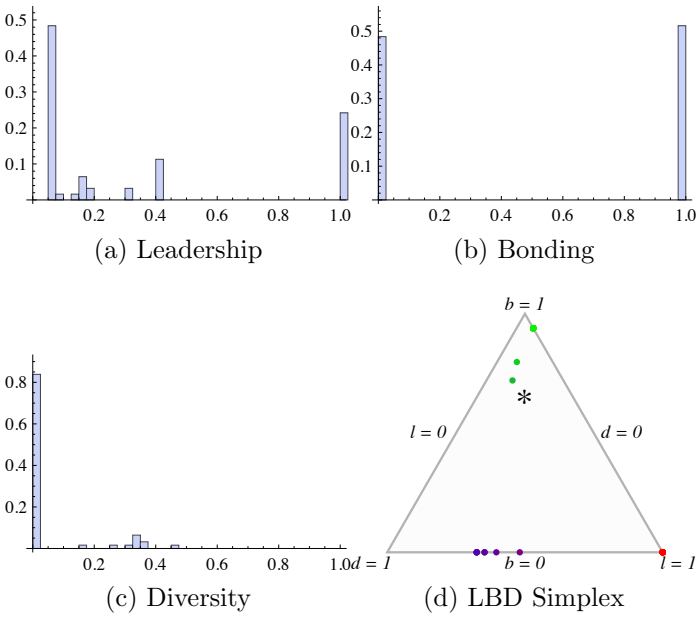


(a) Leadership (b) Bonding

(c) Diversity (d) LBD Simplex

Figure 4-6: LBD distributions and *lbd* simplex for the Tree Clique graph at radius 1.

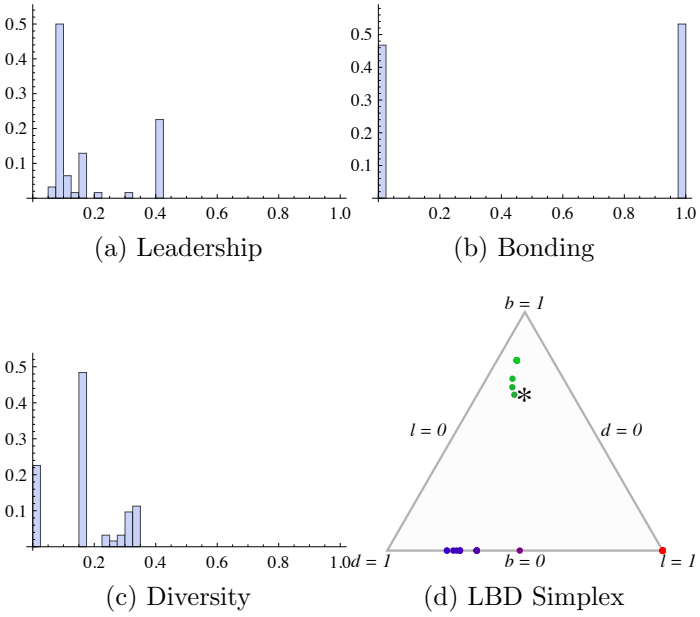
At $r = 1$ there are 4 classes of subgraph. The dipoles at the leaves of the binary tree, which can be seen as the *lbd*-tuples at $(0, 0, 0)$ in the simplex. The neighborhoods of the non-leaf vertices in the tree form stars, seen at $(1, 0, 0)$. Within the complete graph B is maximal, whereas L and D are 0, resulting in the $(0, 1, 0)$ points in the simplex. Finally, the subgraph centered on the vertex within the clique which links to the tree is an outlier, with only one node not having a degree of 31 like the rest of the clique, resulting in an L score of around 0.06, and B score just below 1. This can be seen both in the histograms and the simplex.



(a) Leadership (b) Bonding

(c) Diversity (d) LBD Simplex

Figure 4-7: LBD distributions and *lbd* simplex for the Tree Clique graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) LBD Simplex

Figure 4-8: LBD distributions and *lbd* simplex for the Tree Clique graph at radius 3.

As the local subgraphs expand their radii, the complete graph's vertices connect with the vertices in the tree, seen in the low-yet-increasing L scores. As with the standard binary tree, the subgraphs centered on the leaves at this point form stars, giving them *lbd*-tuple scores of (1,0,0) and the non-leaves come to include other vertices whose degrees are equal to the maximum degree nodes in their subgraph, driving their L scores down. A clear division in B exists between those subgraphs which contain the full complete graph and those which do not. The non-zero D scores belong to subgraphs close to the root of the full tree. In the simplex, 3 neighborhoods can be distinguished: the subgraphs containing the clique in the top of the simplex, the leaves at the bottom right, and the vertices near the root at the mid-bottom.

The 3 neighborhoods are again distinguishable in the scatterplot at $r = 3$. Those centered on the 14 leaves of the tree have $B = 0$, $D = 0$ and L of around 0.4. This places them in the bottom right of the simplex. The subgraphs around vertices near the complete graph are again at the top of the simplex, although descending as their B scores decrease relative to their L and D scores from branching out further into the rest of the tree. The subgraphs not yet in contact with the clique and not centered on the leaves appear at the bottom of the simplex, with their relative D scores increasing as in the standard binary tree case.

4.3.3 Erdős-Rényi graph

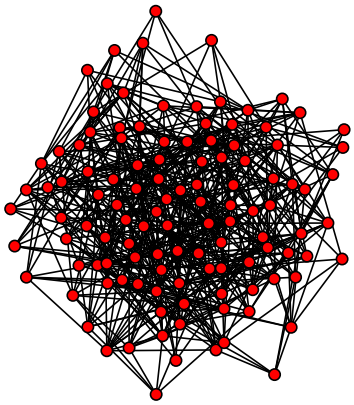
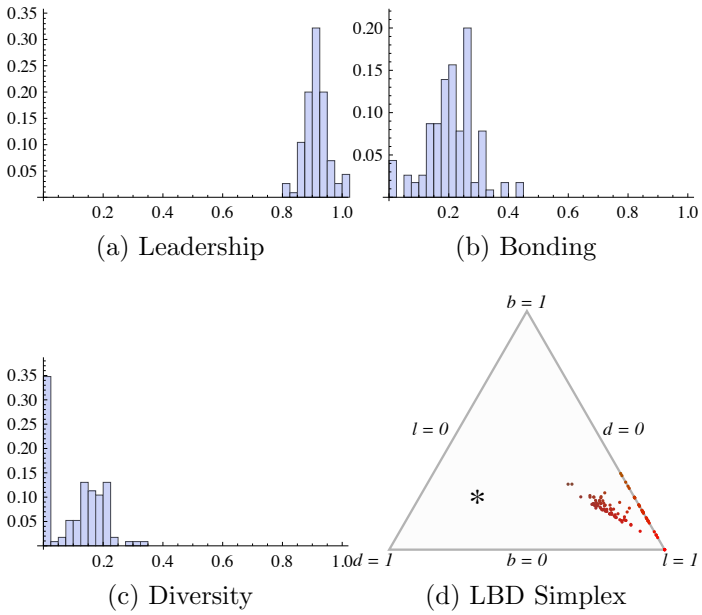


Figure 4-9: The Erdős-Rényi random graph.

As discussed in chapter 2, an Erdős-Rényi random graph is constructed by selecting a graph order, a coin weight, p, and then flipping a coin with that weight for each possible edge position in the graph and adding an edge there if the flip comes up heads. As such, a particular set of parameters defines a whole family of random graphs of which the graph presented here is one example with $|V| = 115$ and $p = 0.1$. These random graphs are well studied and their full graph LBD scores have previously been found empirically to cluster in a region of LBD space occupied by the graph presented here [32]. They tend to be bonded in proportion to their p parameter, since for each 2-path the probability of triadic closure is by definition p [24]. L tends to be low and D relatively high, due to the low probability of disproportionately high degree vertices and the high size to order ratio respectively.



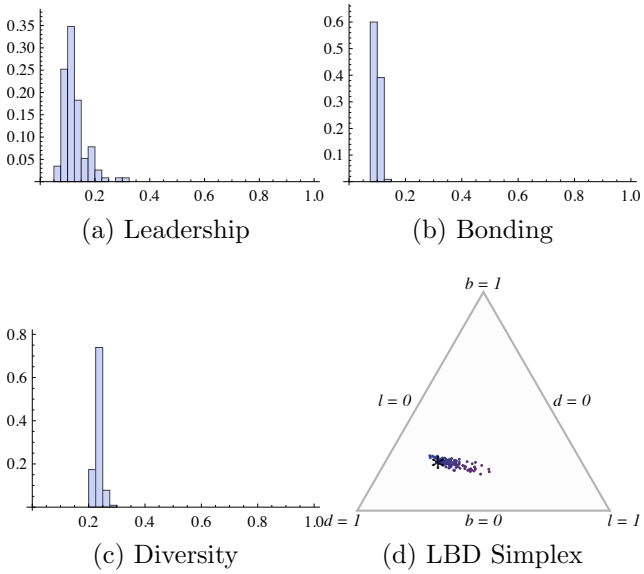
(a) Leadership (b) Bonding

(c) Diversity (d) LBD Simplex

Figure 4-10: LBD distributions and *lbd* simplex for the Erdős-Rényi graph at radius 1.

At $r = 1$ two thirds of the subgraphs have non-zero D scores, which as we will see is quite uncommon. This means that there must a large number of 4-paths in the subgraphs at $r = 1$. The high L scores imply that most of the subgraphs are dominated by one or two vertices with higher degree than those of the rest of the subgraph and from the points at (1,0,0) in the simplex it can be seen that some stars

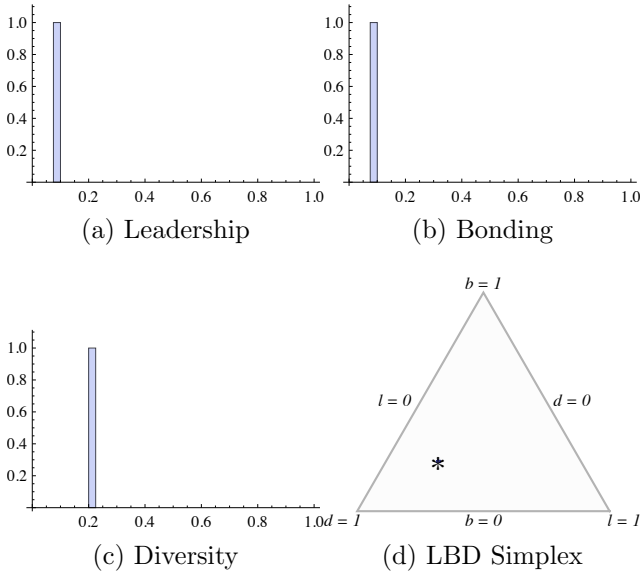
exist. The range of of L and B indicates a heterogenous set of subgraphs.



(a) Leadership (b) Bonding

(c) Diversity (d) LBD Simplex

Figure 4-11: LBD distributions and *lbd* simplex for the Erdős-Rényi graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) LBD Simplex

Figure 4-12: LBD distributions and *lbd* simplex for the Erdős-Rényi graph at radius 3.

Because of the CPL of just over 2 and diameter of 4 the subgraphs converge to common LBD scores very quickly. At $r = 2$ we can see that the majority of subgraphs contain a large proportion of the full graph, and hence have very similar LBD scores. D jumps up as more 4-paths are formed and disjoint sections of the graph are connected, L drops as the dominant vertices are connected and compete with one another for maximum degree, and B stabilizes near the 0.1 score which would be expected given $p = 0.1$.

By $r = 3$ the vast majority of subgraphs contain the full graph, causing the LBD scores to completely converge to the full graph score. The outlier path that gives the graph its diameter of 4 has no effect on the LBD distribution.

4.4 Social graphs

4.4.1 Los Alamos

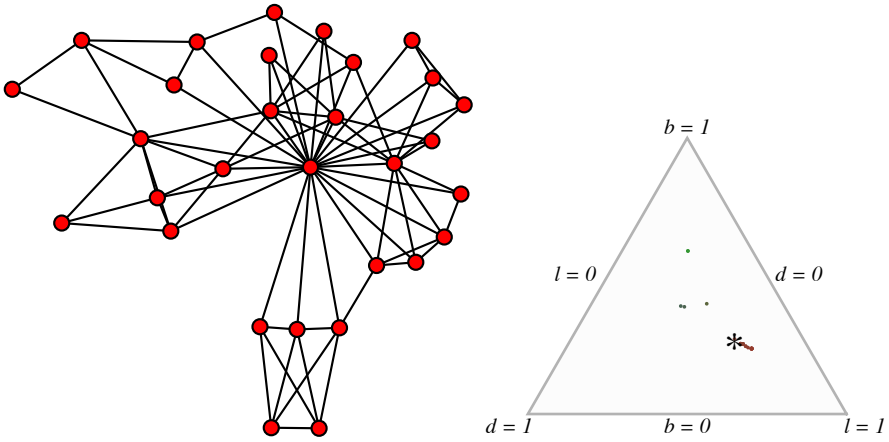


Figure 4-13: Los Alamos graph and $r = 2$ simplex.

The $r = 1$ subgraphs are highly bonded and several cliques exist as is visible in the (0,1,0) points in the simplex. Leadership scores range widely in $r = 1$ neighborhoods, but converge to a high score as subgraph radii increase, dominated by the central node that can be seen in the graph visualization. The low diameter leads to a near convergence to full graph LBD scores by $r = 3$. Even at $r = 2$ the majority of

subgraphs are homogenous, lying close to the full graph *lbd* in the simplex.

4.4.2 Karate

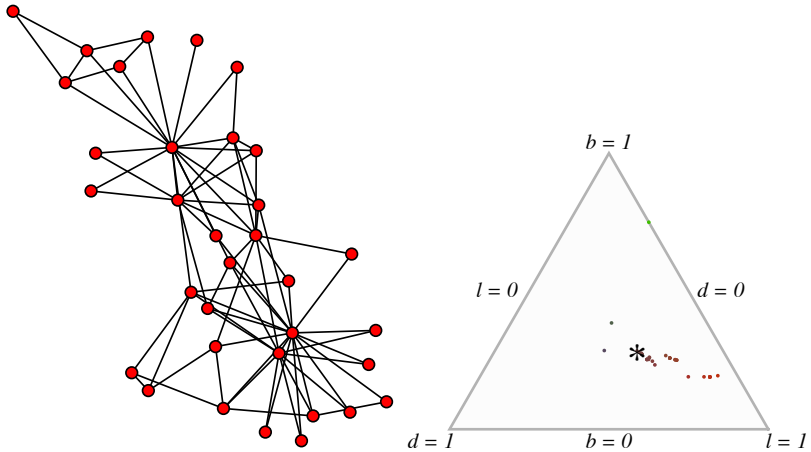


Figure 4-14: Karate graph and $r = 2$ simplex.

The high bonding of $r = 1$ subgraphs shows the existence of cliques, which must be small since their contribution to B is quickly overcome as r increases. A core of highly connected vertices results in a high skewed L distribution at $r = 2$. The low graph diameter and CPL leads to a quick convergence to the full graph LBD scores.

4.4.3 Dolphins

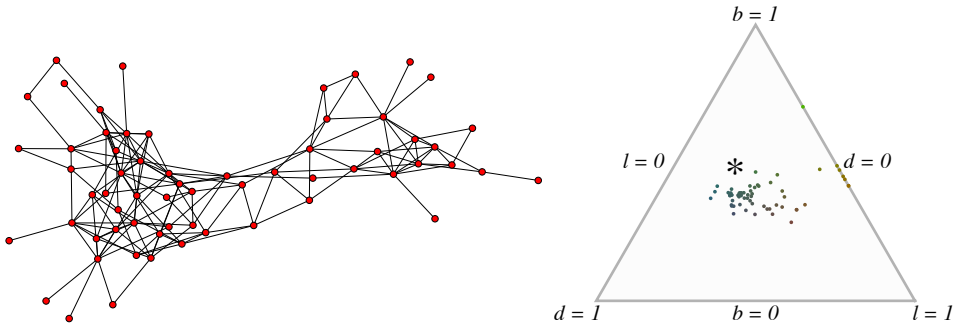


Figure 4-15: Dolphins graph and $r = 2$ simplex.

The high number of (0,0,0) points in the $r = 1$ *lbd* simplex indicates dipoles on

the edge of the graph. The wide range of L and B scores at $r = 1$ and $r = 2$ indicate heterogenous fine structure across the graph. The cloud of points around the center of the *lbd* simplex indicates an even trade off between L, B, and D scores for the $r = 2$ subgraphs.

4.4.4 Enron

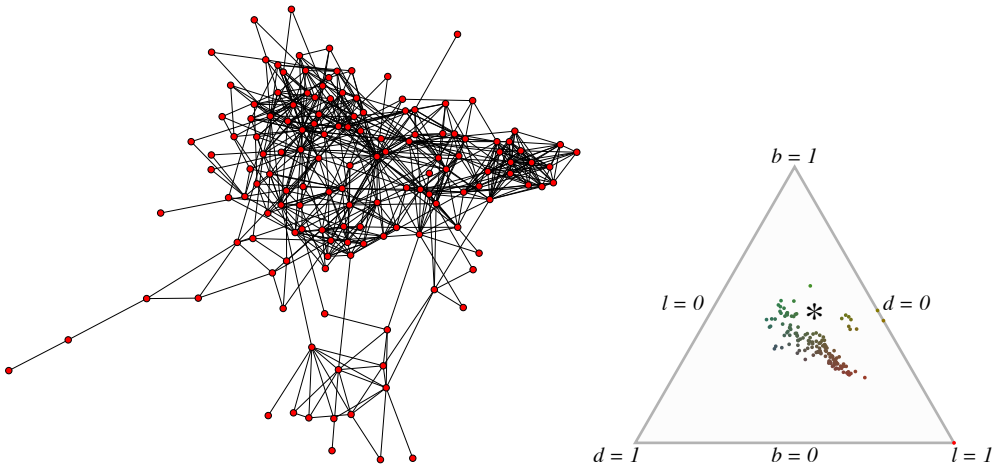


Figure 4-16: Enron graph and $r = 2$ simplex.

Fine structure is again very heterogenous across $r = 1$ and $r = 2$ subgraphs, with a very wide range of LBD scores. At $r = 2$, subgraphs balance B and D scores with L varying widely in proportion as can be seen from the cluster of points across the center of the simplex. At $r = 3$ these L scores begin to converge.

4.4.5 Santa Fe

The broad range of LBD values across all three r values is to be expected for a graph with such a high CPL. The *lbd* scores in the simplex fall within a similar range for both $r = 2$ and $r = 3$, which is likely due to the high CPL causing the subgraphs to only slowly merge with one another. B and D distributions stay relatively constant across radii, whilst L scores drop.

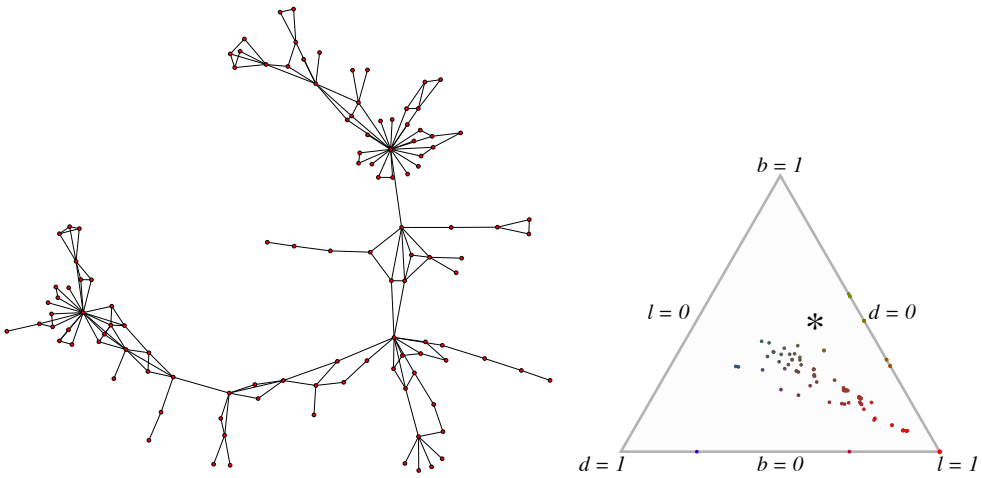


Figure 4-17: Santa Fe graph and $r = 2$ simplex.

4.4.6 JJATT

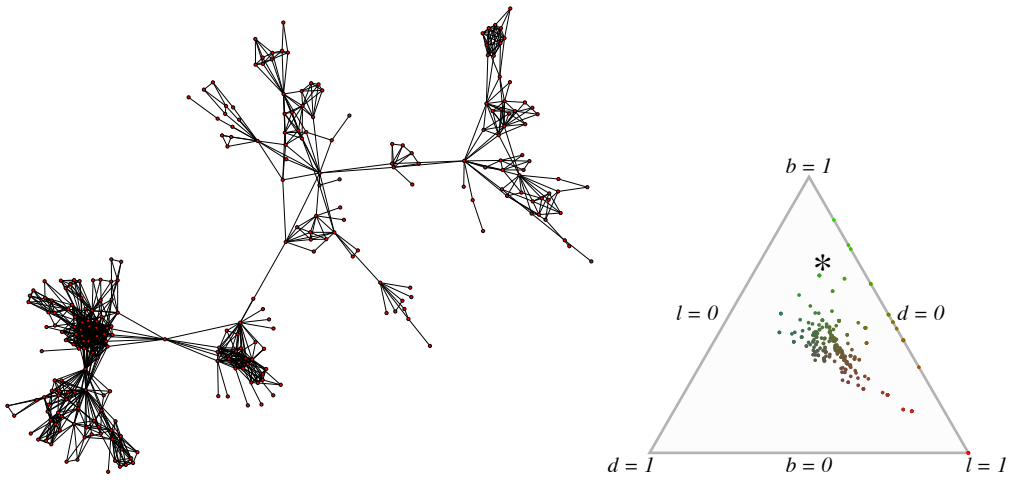
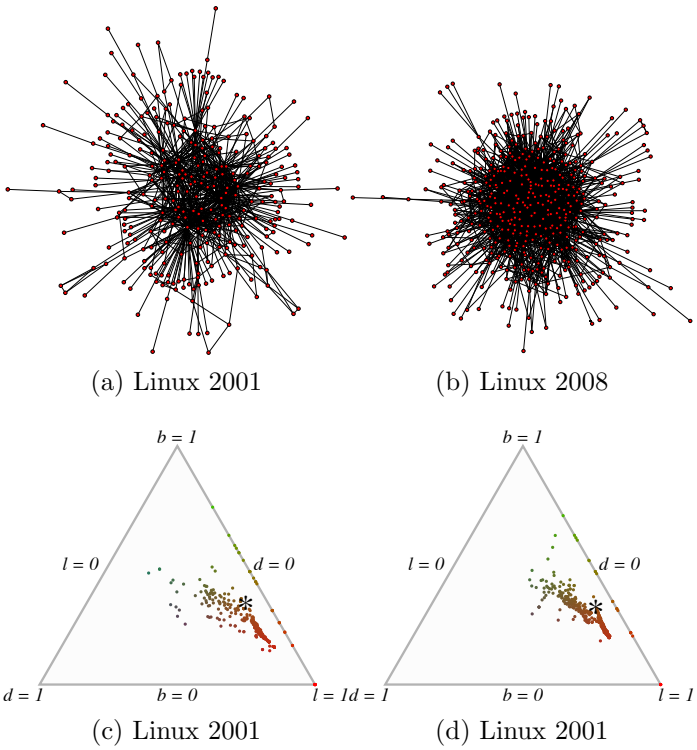


Figure 4-18: JJATT graph and $r = 2$ simplex.

The high CPL of the graph gives it similar convergence properties to Santa Fe, but highly clustered subgraphs result in higher B scores across all three radii. Intuitively this matches what might be expected of the terrorist groups this graph represents. LBD scores vary widely across all three radii, with two modes in L at $r = 2$.



(a) Linux 2001 (b) Linux 2008

(c) Linux 2001 (d) Linux 2001

Figure 4-19: Linux 2001 and 2008 graphs.

4.4.7 Linux 2001 and 2008

The distributions for the two Linux graphs are very similar, with the key difference being the flatter distribution of L scores at $r = 2$ in the 2008 subgraphs compared to the 2001 subgraphs that have higher L scores. This suggests that in 2001 email correspondence was dominated by some key individuals, but over time the load has been spread across several people. In both cases at $r = 1$ and $r = 2$ there are a broad range of subgraph structures, which converge at $r = 3$.

4.5 Other graphs

4.5.1 Bright

Although the $r = 1$ subgraphs have relatively heterogenous structure, the $r = 2$ subgraphs have a narrower range of LBD scores than many of the other graphs in this set, indicating an above average degree of fine structure homogeneity. This is

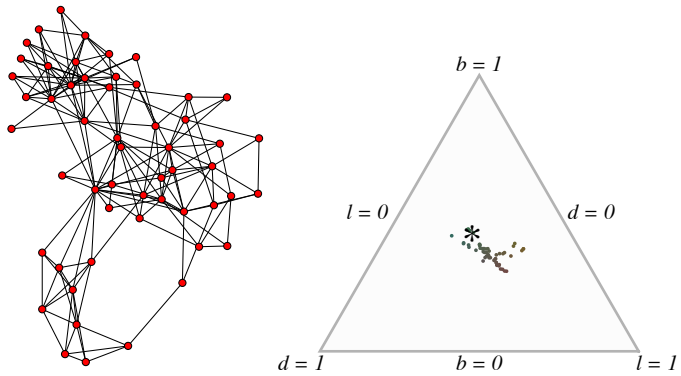


Figure 4-20: Bright graph and $r = 2$ simplex.

also reflected in the $r = 2$ *lbd* simplex by the relatively tight cluster of points around the middle, showing the $r = 2$ subgraphs have a roughly similar trade off between L, B and D scores. By $r = 3$ the low CPL has lead to a near convergences of LBD scores.

4.5.2 Lesmis

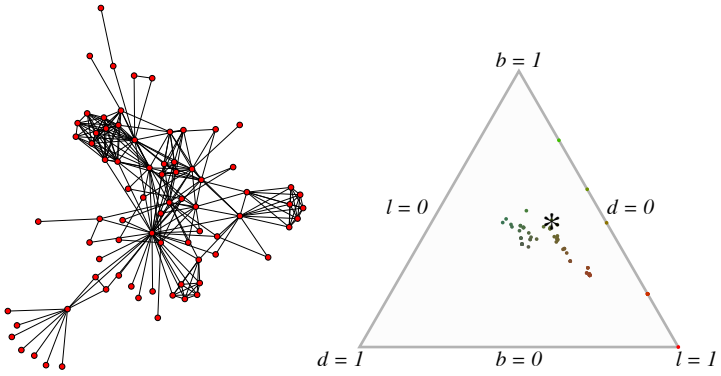


Figure 4-21: Lesmis graph and $r = 2$ simplex.

A broad range of LBD scores exist at $r = 1$ with a significant number of dipoles and several cliques, visible from the $(0, 0, 0)$ and $(0, 1, 0)$ scores in the simplex and the zeros in for all three scores along with the ones for B respectively in the distributions. At $r = 2$, three different characteristic levels of B can been seen in the distributions, as well as two modes for L and D, which manifests as three distinct clusters of points

in the simplex. These heterogenous regions have largely merged at $r = 3$, but a spread of L scores can still be seen.

4.5.3 PolBooks

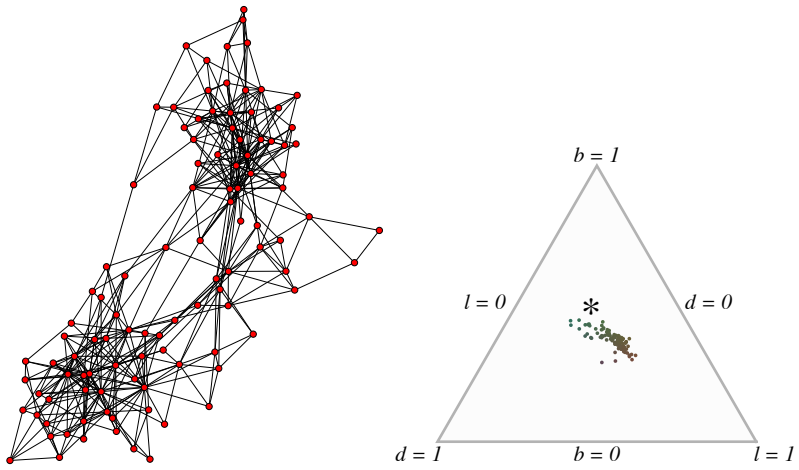


Figure 4-22: PolBooks graph and $r = 2$ simplex.

A heterogenous $r = 1$ LBD distribution but a relatively homogenous $r = 2$ distribution can be seen in this graph, with B and D distributions having dominant modes and L varying the most. This is reflected in the tight clustering of points in the $r = 2$ simplex around the L axis, which becomes even tighter at $r = 3$ as the subgraphs converge towards the full graph LBD score.

4.5.4 AdjNoun

At all three radii the subgraphs of this graph show L scores distinctly more dominant than B and D. At $r = 2$, B and D have distinctive low modes, whilst L has a very wide spread, including some markedly high scores. At $r = 3$ the L scores of the subgraphs converge to the full graph's lower LBD score, but dominance of L in the subgraphs relative to B and D is clearly visible in the simplex.

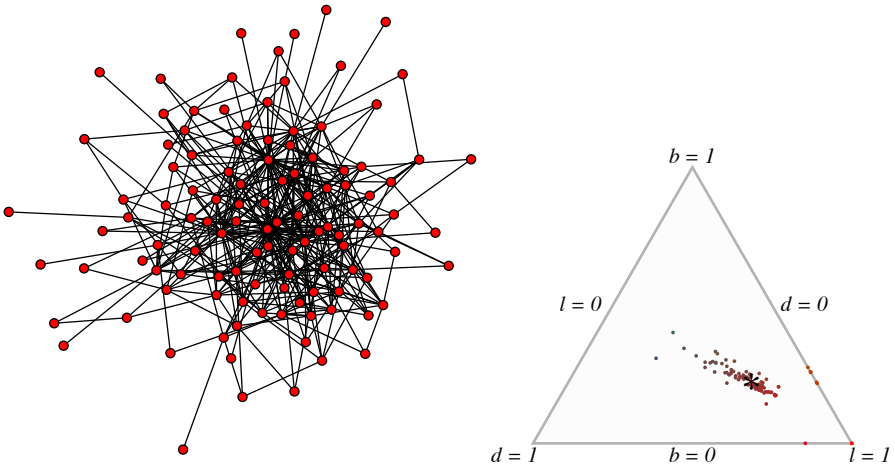


Figure 4-23: AdjNoun graph and $r = 2$ simplex.

4.5.5 Football

This graph is particularly interesting due to its dramatic shift between $r = 1$ and $r = 2$. At $r = 1$, the subgraphs have a spread of relatively high leadership scores and high bonding. This is due the the structure of the Football competition in the graph, where many of the subgraphs are groups of teams playing each other in local competition. Some vertices of these subgraphs are victors and therefore play more games, leading to the high L scores. At $r = 2$ the victors play each other, so their high degrees cancel each other out and lead to the drop in L scores visible in the L

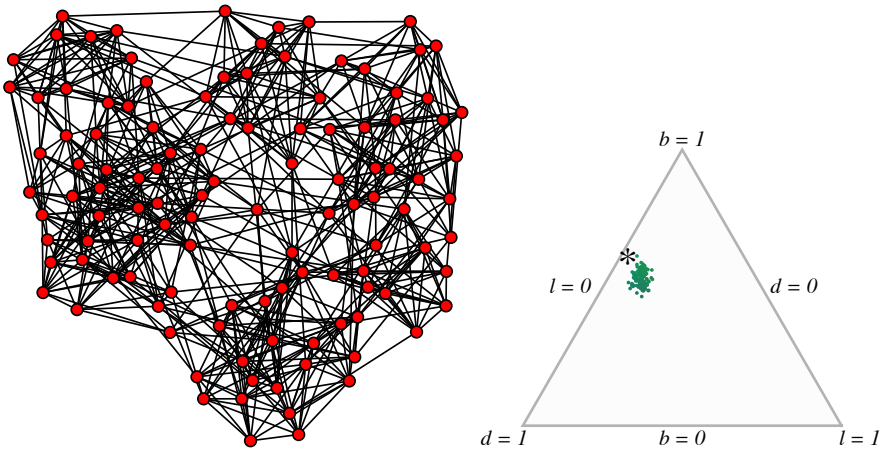


Figure 4-24: Football graph and $r = 2$ simplex.

distribution. At $r = 3$ most subgraphs contain the full graph and converge to the full graph LBD score.

4.5.6 C-Elegans

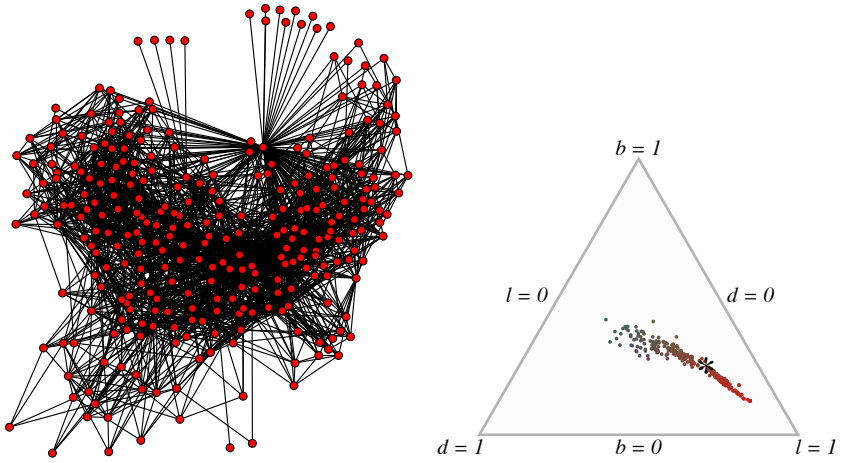


Figure 4-25: C-Elegans graph and $r = 2$ simplex.

The subgraphs at $r = 1$ in this graph have a very heterogenous structure, with LBD scores ranging very widely, but at $r = 2$ the majority of the variance in these scores is in their L values, with B and D having very distinctive modes. This spread of L scores dramatically converges however at $r = 3$.

4.5.7 PolBlogs

This graph contains isolated vertices, cropped from the visualization, which can be seen in the $(0, 0, 0)$ scores in the center of the simplex at all three radii. The subgraphs are heterogenous at $r = 1$ and $r = 2$ in all three dimensions, particularly L and B, which show wide variation. At $r = 3$ however, the LBD scores of the subgraphs converge to a relatively small region in spite of the graph's high CPL.

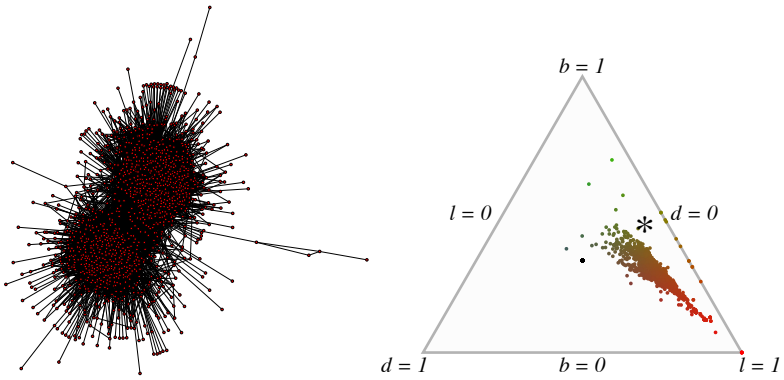


Figure 4-26: PolBlogs graph and $r = 2$ simplex.

4.6 Discussion

These results show clearly that graphs with similar full graph LBD scores can have quite different fine structure. Consider AdjNoun and PolBooks, both of which have full graph LBD scores in a similar range. At $r = 1$ these graphs are widely different, with AdjNoun containing many more stars as a proportion of its subgraphs than PolBooks. Alternatively consider two graphs in a similar region of the *lbd* simplex such as Santa Fe and Lesmis. The distribution analysis shows that the LBD scores, and therefore the structure, of their local subgraphs are widely different, despite their full graph LBD scores having similar relative proportions. The LBD distribution representation therefore fills a gap in the full graph LBD representation of structure.

For the majority of graphs analyzed, the scale of the analysis, r, has a strong impact on the distribution of LBD values observed and by extension the graph's fine structure. The clear exception is for graphs with high CPLs such as Santa Fe, in which increasing the radius of the subgraphs analyzed introduces relatively few new vertices and edges, resulting in only gradual changes in fine structure. For every graph there is a dramatic shift in subgraph structure between $r = 1$ and $r = 2$, due to including many more edges that can contribute to higher D scores and make stars, complete graphs, and isolated dipoles far less prevalent structures. In this sense $r = 2$ is more informative about graph fine structure than $r = 1$, preserving information about small neighborhoods, whilst being a large enough radius to let LBD scores

respond to connections between linked neighborhoods.

Chapter 5

Graph similarity and clustering

5.1 Motivation

Since the LBD distribution of a graph summarizes its fine structure, we can compare the LBD distributions of two graphs to judge their similarity. Visualizations of LBD distributions can be qualitatively compared to give a broad coarse grained feel for the similarity of two graphs, but visualizing the joint distribution is challenging due to its 4 dimensional structure (L, B, D, and frequency), leaving us to rely on the strategies such as the one presented in the previous chapter, inspecting distributions for each dimension separately and projecting the data into a lower dimensional space. A quantitative comparison of LBD distributions is computationally feasible however. By computing the earth mover's distance [30, 27, 33] between two LBD distributions a single value can be derived that summarizes the similarity of fine structure between two graphs. In this chapter I demonstrate this method and use it to perform agglomerative graph clustering, building a hierarchy of graphs based on the similarity of their fine structure.

5.2 Method

In performing a fine structure comparison there are some choices and tradeoffs to be made. One is what subgraph radius to consider for the distributions. In the

extreme case one could compare two graphs where the radius of the comparison is the diameter of one graph, but much smaller than the diameter of the other. This would reveal little interesting about the relationship between the two. For much social network analysis, researchers are interested in ego-centric subgraphs with 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 [38]. As discussed in the previous chapter, radius 2 is a reasonable compromise between preserving information about small neighborhood structure, whilst responding to connections between linked neighborhoods.

An issue which was not mentioned in chapter 4 is whether or not to make the LBD space discrete when computing distributions. Previously LBD distributions were discretized into histogram bins without much thought for how large those bins ought to be. The choice of the granularity of this discretization will impact any comparison, since coarser discretizations may place distinct LBD scores in the same bin. For this thesis I 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. The 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. The construction of LBD distributions presented in the previous chapter 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 could be simply a vector of each LBD value 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 are being compared and one is a subgraph of the other, much 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

the distributions are normalized, a comparison between two graphs of different sizes is possible, whereas in a presence-based representation this may add complications.

I perform a 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. I then normalize the counts of the histogram bins by dividing by the number of vertices in each graph, yielding two LBD distributions. I compute the earth mover's distance between these two distributions using Euclidean distance as the ground distance. Finally I normalize by the maximum distance in the discretized space and subtract the result from 1 to yield a similarity measure in the range $[0, 1]$.

Armed with a method for judging graph similarity by fine structure features, I use it to find classes of graph that have these features in common using a clustering approach. There are many choices of clustering algorithm available, so I have opted for the generality and simplicity using average-link hierarchical clustering following the method in [9]. In this agglomerative approach to clustering one computes the pairwise similarities of all the graphs in the set to be clustered. Initially each graph is in its own cluster. At each step ones then merges the two clusters for whom the mean similarity is highest, resulting in a hierarchy of graph clusters.

5.3 Similarity results

To demonstrate that this similarity measure produces intuitively plausible results, I followed the example of [29] and computed the similarity of a representative set of the graphs from the previous chapter to permutations of themselves. I parameterized these permutations by a noise factor ranging in 0.1 increments from 0 to 0.8 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 5-1, which demonstrates, as hoped, that this similarity measure judges graphs to be less similar to their permutations as the degree of permutation increases. Note however the case of the Erdős-Rényi random graph. This is the top line in the plot, almost coincident with the top of the figure. The consistent high

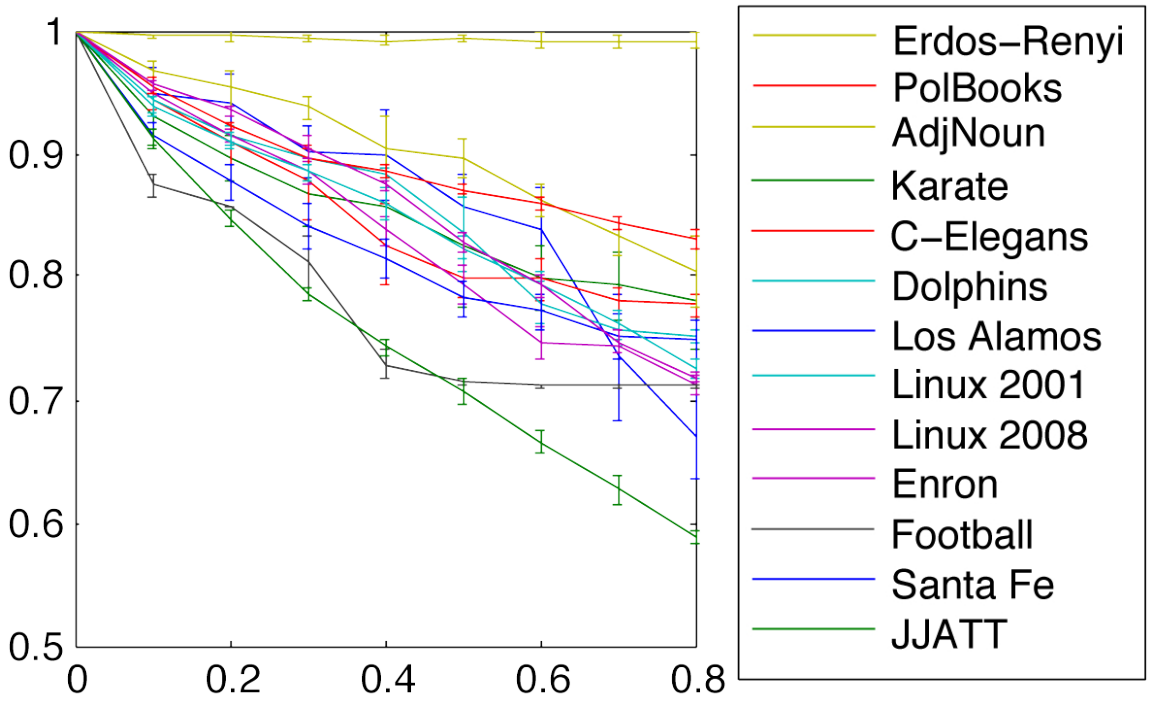


Figure 5-1: Radius 2 self-similarity. Ordinate: fraction of edges permuted; abscissa: earth mover's distance similarity measure.

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 leads them to have characteristic fine structure properties at $r = 2$, namely low leadership, bonding, and 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.

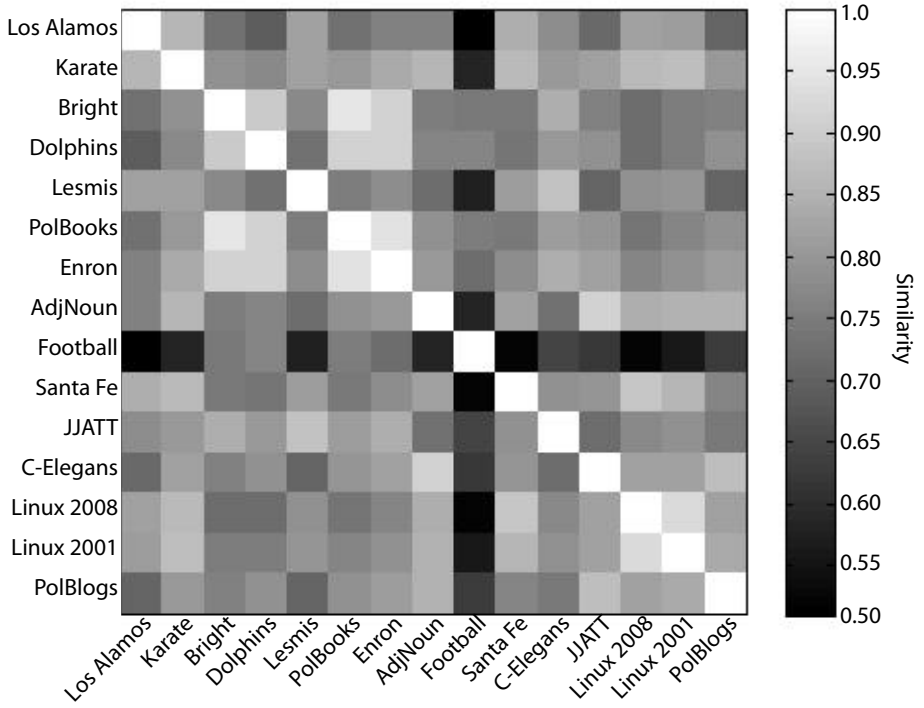
Figure 5-2a shows the pairwise similarity between each graph in the social and other graph sets computed with $r = 2$. By contrast, 5-2b shows the similarity between the graphs judged by the inverse of normalized distance between their full graph LBD scores. From these results it is clear there is a 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 discussed in the previous chapter. This quantitatively confirms the general conclusion drawn at the end of the previous chapter, namely that two graphs can have a similar global structure, judged by their full graph LBD scores, and yet have quite dissimilar fine structures.

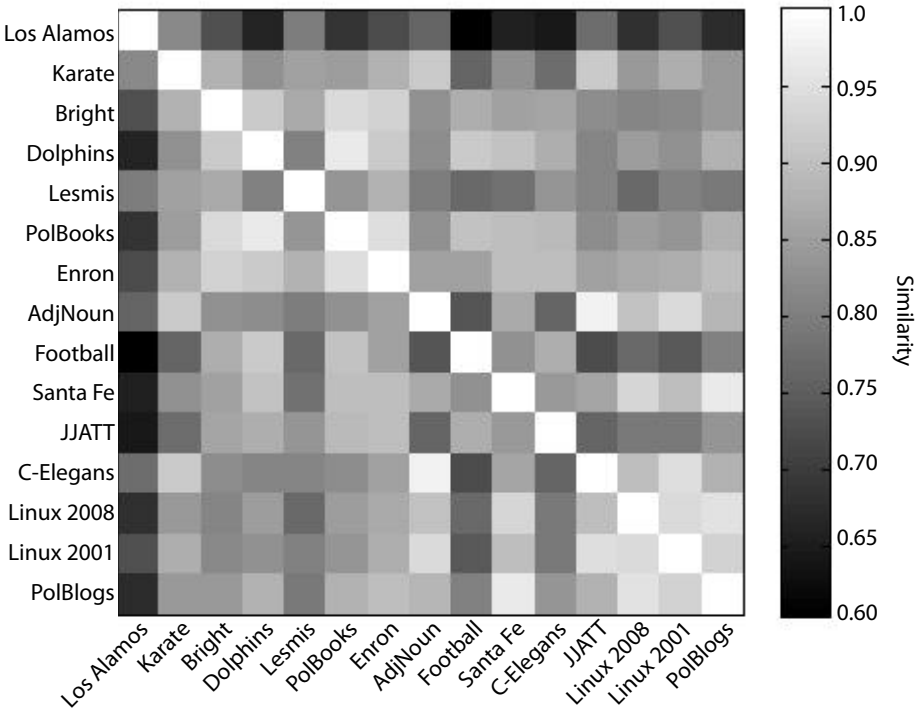
5.4 Clustering results

Figures 5-3a and b show dendrograms for the results of the clustering using the radius 2 and full graph similarties respectively. Recall that each graph is initially assigned to a cluster by itself and these clusters are merged in order of greatest average similarity between cluster members. Clusters are represented by horizontal lines and their merging is represented by vertical lines connecting the lines of the merged clusters. The horizontal axis gives the similarity value at which the merging took place. The names of the graphs drawn from social data are shown in red. Again, a key point is that the results are different, showing that similarity in fine structure and full graph structure are not equivalent. As an illustrative example of this, consider C-Elegans and Linux 2008. The full graph clustering considers these two graphs to be similar, placing them in the same cluster at a high similarity threshold of 0.94, but the fine structure clustering does not, only clustering them at a threshold of 0.80. Their full graph LBD scores are close, but looking at their $r = 2$ LBD distributions, the subgraphs of Linux 2008 have a more restricted range of L scores and a broader range of B scores, leading Linux 2008 to be clustered with Linux 2001 over C-Elegans.

Looking at the clusters formed by the fine structure analysis it is interesting to note that they often contain a graphs from a mix of different domains, 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. AdjNoun, a semantic network, and C-Elegans, a neural network, are the only



(a) Radius 2 graph similarities



(b) Full graph similarities

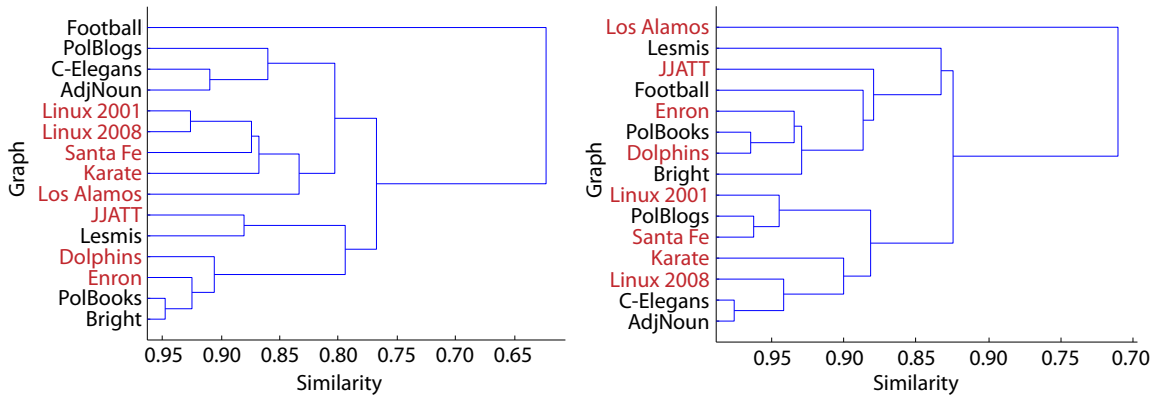
Figure 5-2: Radius 2 and full graph similarities.

two graphs that are judged as being more structurally similar to each other than to any other graphs in the set in both the full graph and fine structure analyses. This judgement stems from the fact that in both cases the LBD distributions of the $r = 2$ subgraphs of 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 structure. Although the fine structure analysis places them in the same cluster hierarchy, they are still judged to be significantly different. 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,



(a) Radius 2 dendrogram

(b) Full graph dendrogram

Figure 5-3: Hierarchical clustering dendrograms based on radius 2 and full graph similarity. Vertical connections show the similarity at which clusters are merged in the hierarchy. The names of graphs generated from social data are shown in red.

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 not from 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.

Chapter 6

Conclusions and future work

6.1 Conclusions

In chapter 4 I introduced LBD distributions as a representation of fine grained structure in graphs. Using this representation to analyze a variety of graphs revealed that graphs could have similar full graph LBD scores or have LBD scores that stood in the same ratios to one another, and yet have fine structures that are widely different. This means the LBD distribution representation reveals more about the structure of a graph than is apparent by looking at its macro level features. Furthermore I showed that for some graphs the scale at which fine structure is considered matters a lot for their LBD distribution. The range of LBD values in almost all graphs changes dramatically between $r = 1$ and $r = 2$ due to the increasing role of D and the reduction in the number of stars and isolated dipoles. I argued that $r = 2$ is a fruitful scale for analysis, because it is the smallest radius that is still large enough to detect structural features such as bridges between communities in a social network. I also analyzed a number of graphs for whom the LBD scores of their subgraphs converges only slowly to the full graph LBD score, namely graphs with large relative CPLs. This empirical result makes sense when we consider that for graphs with high CPLs we should expect that each increase in radius causes the subgraphs being analyzed to only modestly increase in size, and therefore we shouldn't expect changes in their LBD scores to be as dramatic as for graphs with lower CPLs.

I demonstrated in chapter 5 that the earth mover's distance between two LBD distributions is a reasonable approximation of edit distance, and therefore graph similarity, by comparing graphs against permutations of themselves. An interesting side result was that this was not true for Erdős-Rényi random graphs and I argued that the reason for this was that edge permutation merely transforms one such random graph into another instance of a random graph. Since a subgraph of such a random graph is itself random, and it has previously been demonstrated empirically that such graphs have a characteristic range of LBD values, we should not be surprised that these $r = 2$ subgraphs likewise have a characteristic range of LBD values.

One might naively think that there could be fine structure features that characterize networks of social interactions generally, or some class of social interactions such as email exchanges specifically. The results of chapter 5 provide some weak evidence to support this idea, but the results are inconclusive. Whether this is even a reasonable expectation is unclear, given that the processes represented by edges and vertices in graphs of social data can be quite different. Even when two graphs purport to represent the same attributes, such as a friendship tie, the methods by which they are constructed can be quite different, with implications for the resulting graph structure.

With these caveats in mind, I nevertheless demonstrated that average-link hierarchical clustering on $r = 2$ LBD distribution similarity forms a homogenous cluster of graphs derived from social data, but this result was weakened by the fact that several graphs derived from social data fell outside of the cluster, in particular the Enron graph. The Enron and Linux email correspondence graphs were judged to have widely different fine structures, which argues against there being a characteristic fine structure for email networks across institutions. However, the fact that the two Linux email networks were judged highly similar suggests that even if email correspondence data in general lacks a characteristic fine structure, it may still be the case that the correspondence data within institutions can have a characteristic fine structure that is largely time invariant.

By contrasting the clustering results based on fine structure with those based on

full graph structure I demonstrated that there are significant differences in judgements of similarity depending on the granularity of the structural comparison. This provided quantitative evidence for the qualitative argument that I posed in chapter 4, namely that analysis of fine structure is not redundant in the face of information about full graph structure; it reveals structural information that is not merely equivalent to that derived from a full graph analysis. I also pointed out that clustering based on fine structure created clusters that were more homogenous than those based on full graph structure with respect to whether or not their network data came from social sources.

6.2 Future work

6.2.1 Graph structure

Some of the graph data analyzed in this thesis was originally directed, but was converted to undirected graph data in order to be able to measure L, B, and D. This transformation results in a loss of information about graph structure and can lead to counterintuitive results. For instance, a graph may be judged to be highly bonded when in fact many nodes within the graph are unreachable from others along the directed edges. A natural extension of this work would be to develop versions of L, B, and D which take edge direction into account. These directed versions could then be used to generate higher fidelity LBD distributions, which better reflect the directed graph structure.

Another direction could be to take the idea of characterizing graph structure by the features of subgraphs and apply it with a completely different set of structural features. In this thesis I chose L, B, and D because of their social dimension, but one could imagine features measuring other structural dimensions such as reachability or centrality. The general framework of decomposing a graph into subgraphs with a chosen radius, measuring the distribution of the structural features of those graphs, and then comparing those distributions between graphs is easily extensible in this way.

The fact that Erdős-Rényi random graphs have empirically been found to lie in a relatively constrained region of LBD space has been noted both in this thesis and previously, but as yet there is no formal analysis of why this observation should be so. The fact that, as shown in this thesis through the results on permutation of random graphs, such graphs additionally have characteristic LBD distributions is yet more motivation for a formal analysis of the structural properties of random graphs with respect to L, B, and D.

Throughout this thesis I have visualized LBD distribution data by presenting the distribution of values for each dimension independently. This was a compromise because of the difficulty of visualizing the distribution of joint LBD values. Presenting a point cloud of the ratios of these values in the *lbd* simplex went some way to giving a qualitative feel for this data, but it still remains challenging for humans to process. A more compelling visualization method would make LBD distributions more informative and useful.

6.2.2 Graph similarity and clustering

In chapter 5 I identified clusters of graphs with similar features in their fine structure. The natural question then is what specifically is this common structure and how does it 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 similarity being tied to a graph's generative process comes from the result obtained for Erdős-Rényi random graphs, where the random edge permutation transformations did not significantly impact the similarity of the original graph to its transformation. A key direction for future research is a statistical analysis of the outcomes of different generative processes for graphs in terms of their LBD distributions.

The fine structure comparison in chapter 5 gave evidence that the process modeled by graphs representing the same phenomena, for instance email correspondence graphs, can lead to quite idiosyncratic graph structure. One might expect that if the Enron and Linux correspondence graphs model a similar generating process, then

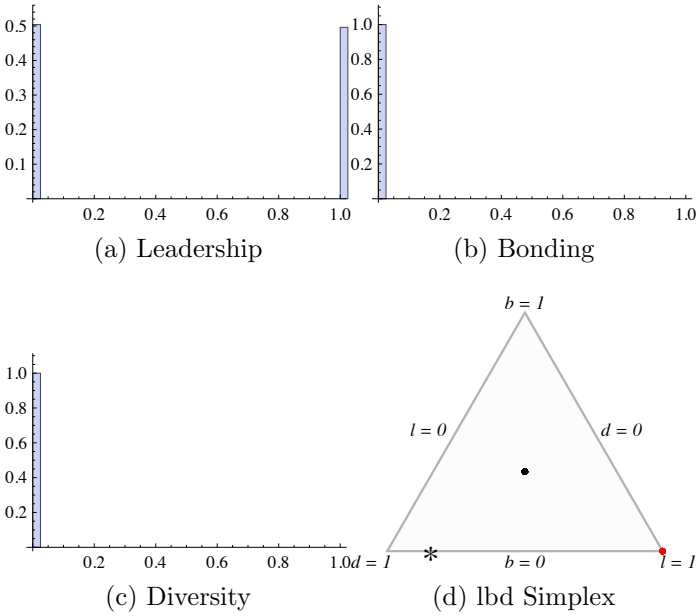
their fine structure should be similar too, but in fact neither their fine structure nor their full graph structure is similar, which suggests that dissimilarities in the organizational structures of Enron and the Linux kernel developers are more crucial factors in the formation of the graphs than the mere fact that the graphs represent email correspondence. The specific nature of these idiosyncrasies and what they reveal about the efficiency, centralization, or other characteristics of the underlying organizations that produce the data in these graphs is a topic for future work.

Despite finding dissimilar fine structure between email graphs, clustering on fine structure revealed a cluster containing the majority of the graphs in the set derived from social data. Although a weak result, it bears consideration that automatic classification of graphs into social and non-social could potentially be done using fine structure, assuming that this similarity result holds beyond the set of graphs presented in this thesis. Less speculatively, the similarity of the two Linux graphs suggests that it may be possible to automatically distinguish distractor graphs from graphs of social data, such as email correspondence, generated by the same organization at different points in time.

Appendix A

LBD distribution figures

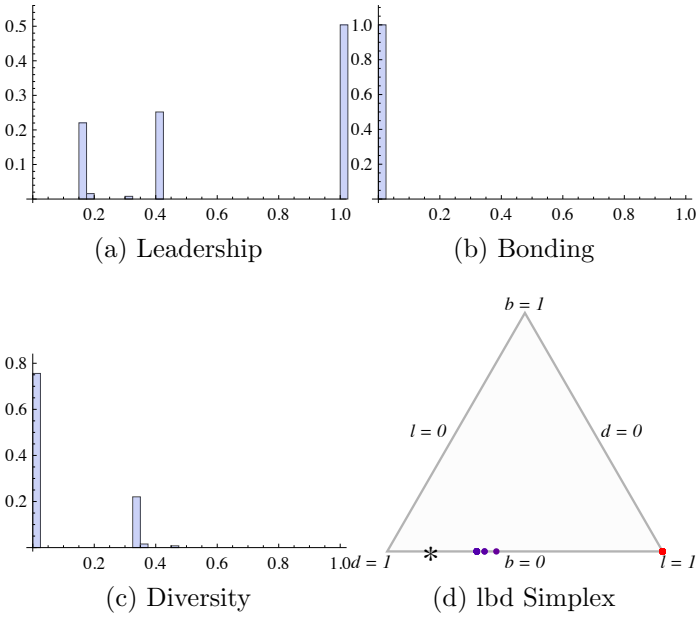
This appendix presents L, B, and D distributions at radius 1 through to 3 for each graph studied in this thesis, along with simplex visualizations showing the relative magnitudes of each of those scores to one another for each subgraph. In each simplex visualization the full graph's *lbd* location is represented with an asterisk as a point of reference. I have made the data for the Bright, Enron, JJATT, Linux 2001, Linux 2008, Los Alamos, and Santa Fe graphs available on the web at `http://people.csail.mit.edu/owenm/netdata.html`. The remainder of the graphs appearing in this thesis can be found in Mark Newman's collection of network data at `http://www-personal.umich.edu/~mejn/netdata/`.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

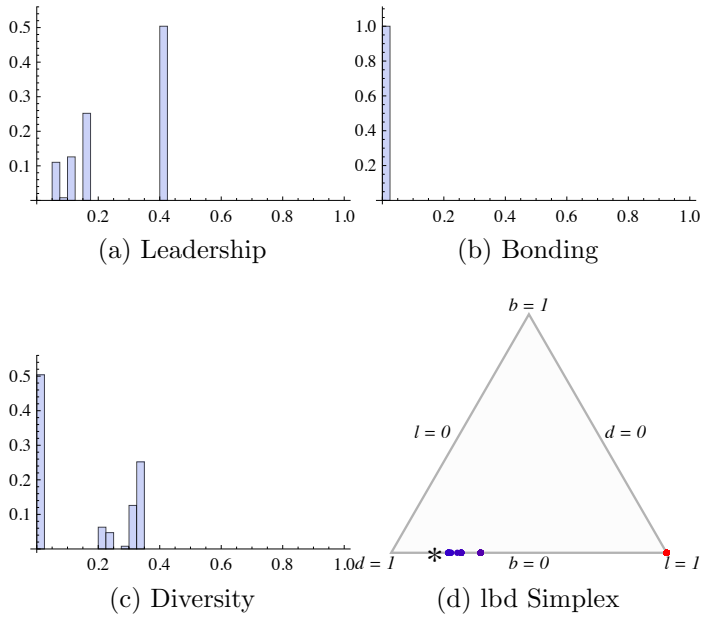
Figure A-1: LBD distributions and lbd simplex for the Binary Tree graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

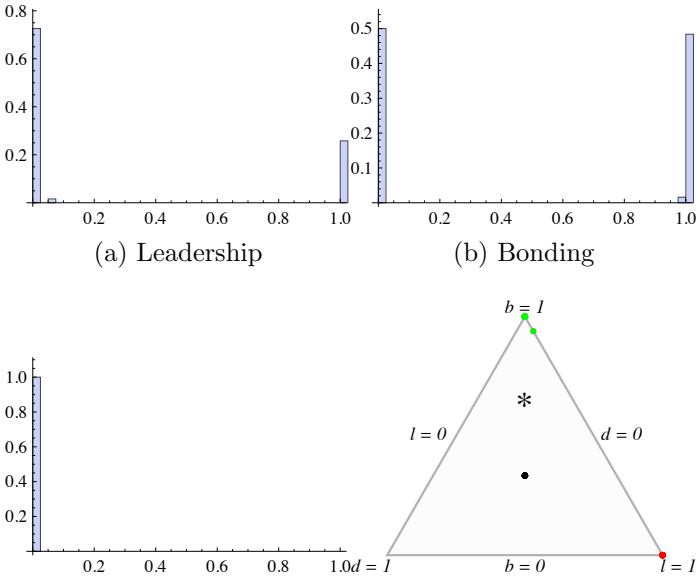
Figure A-2: LBD distributions and lbd simplex for the Binary Tree graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

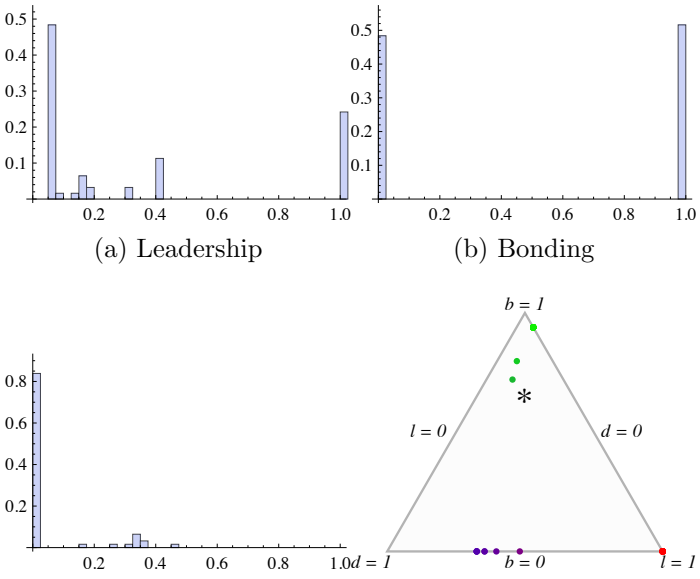
Figure A-3: LBD distributions and *lbd* simplex for the Binary Tree graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

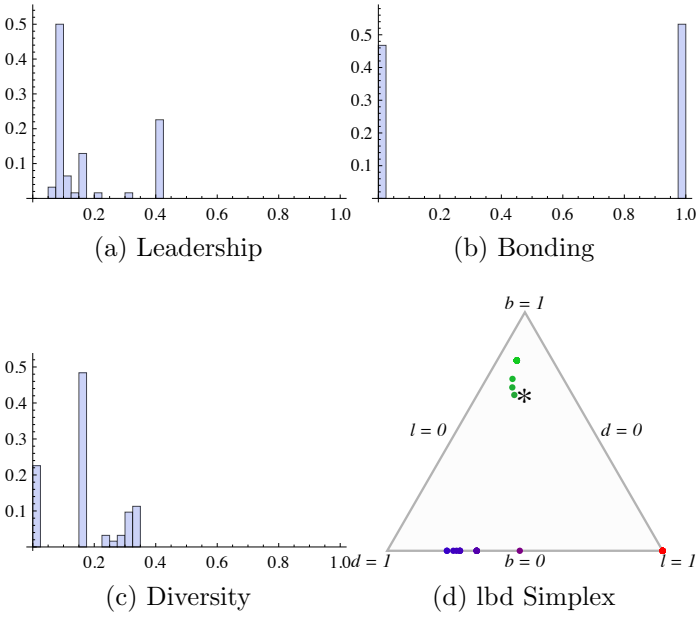
Figure A-4: *LBD* distributions and *lbd* simplex for the Tree Clique graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

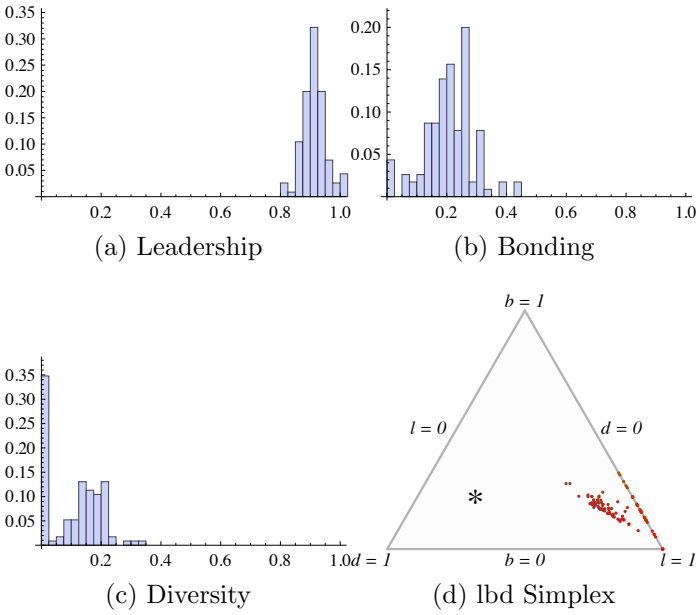
Figure A-5: *LBD* distributions and *lbd* simplex for the Tree Clique graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

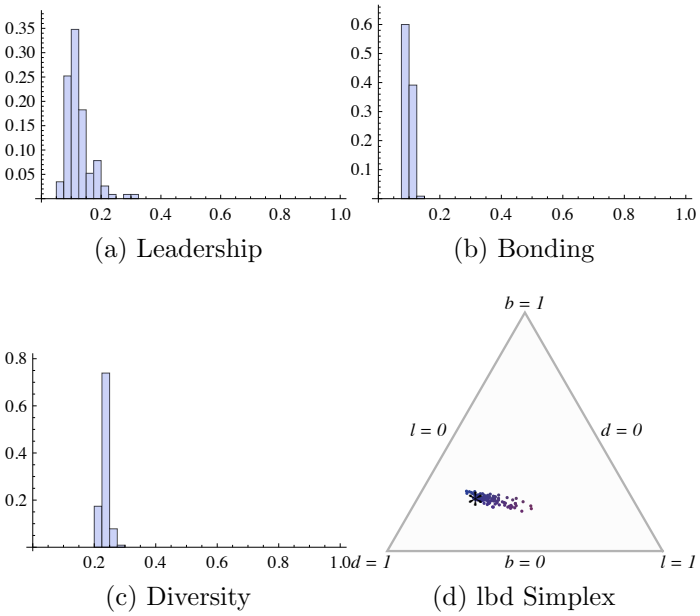
Figure A-6: LBD distributions and *lbd* simplex for the Tree Clique graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

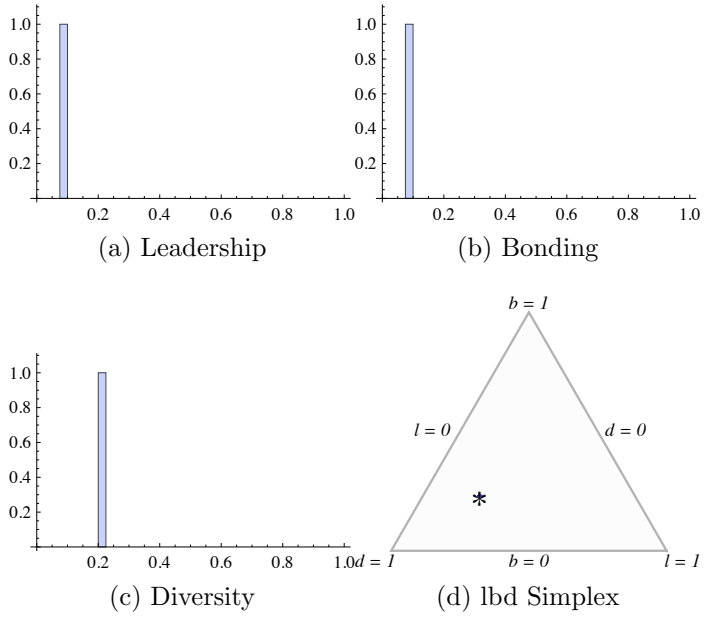
Figure A-7: *LBD* distributions and *lbd* simplex for the Erdős-Rényi graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

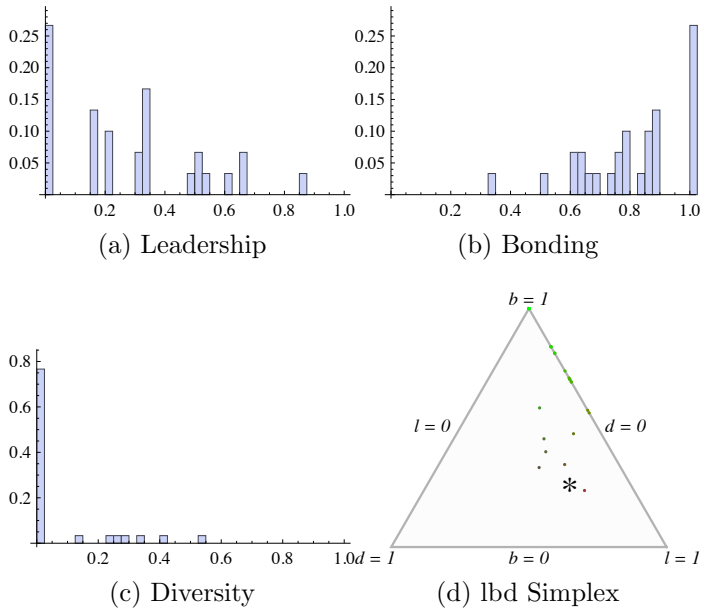
Figure A-8: *LBD* distributions and *lbd* simplex for the Erdős-Rényi graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

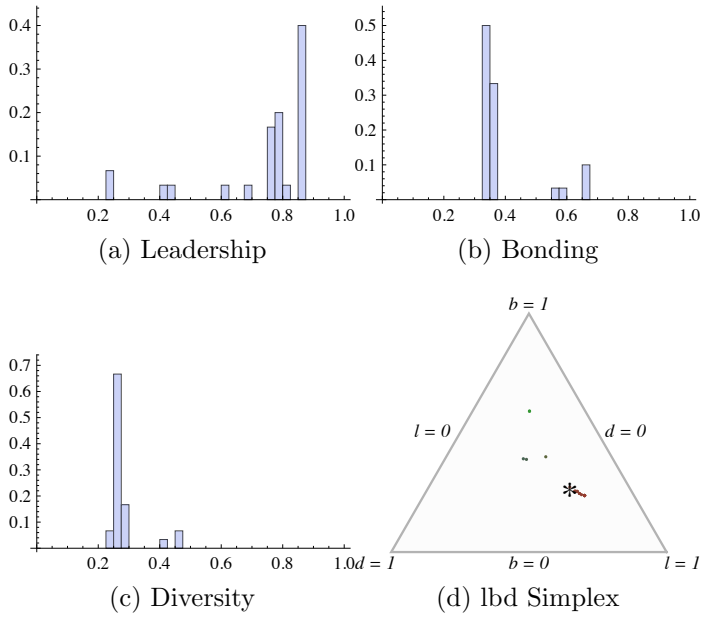
Figure A-9: LBD distributions and *lbd* simplex for the Erdős-Rényi graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

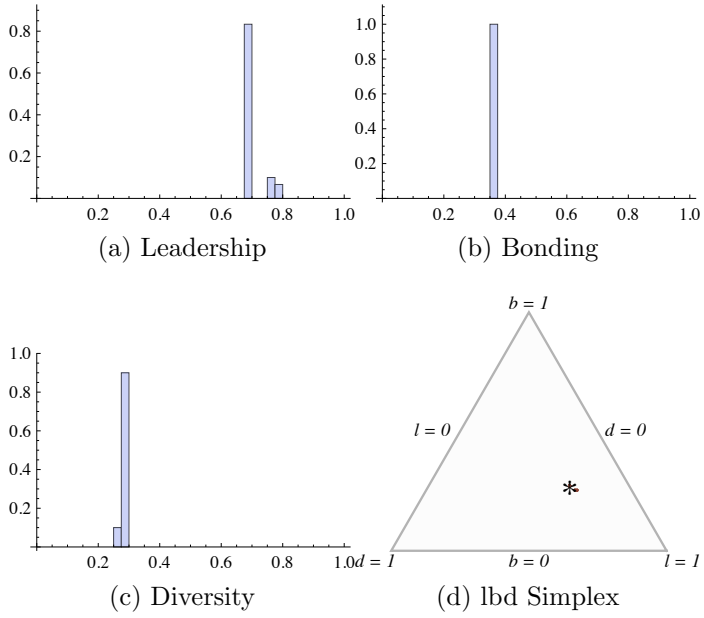
Figure A-10: LBD distributions and lbd simplex for the Los Alamos graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

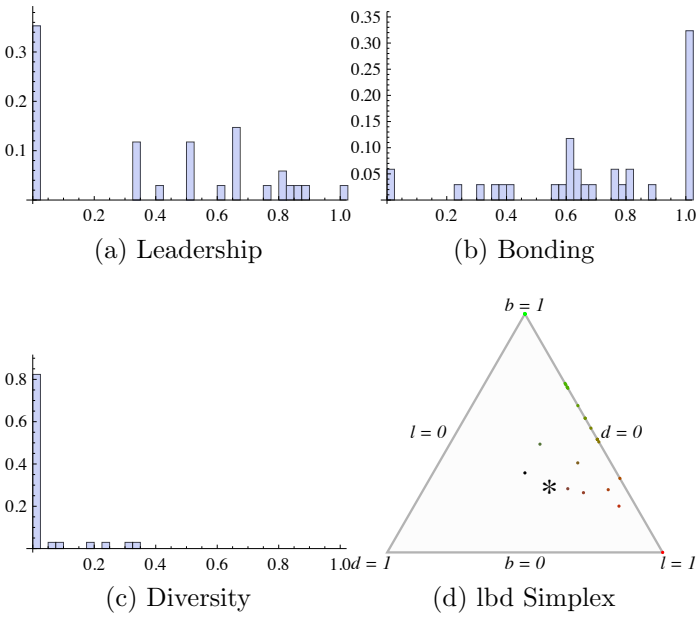
Figure A-11: LBD distributions and lbd simplex for the Los Alamos graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

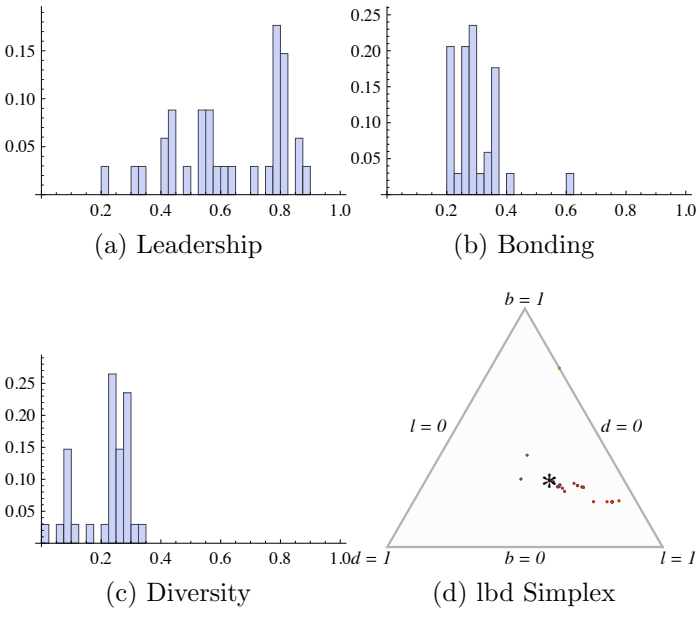
Figure A-12: LBD distributions and *lbd* simplex for the Los Alamos graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

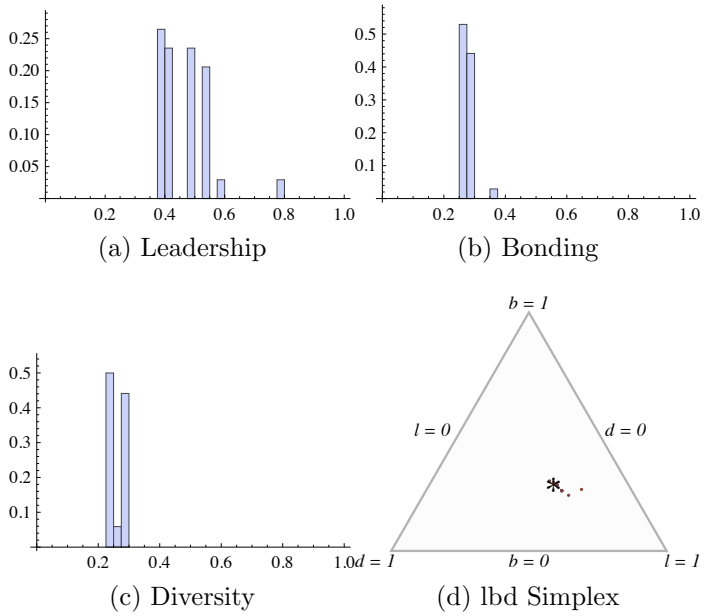
Figure A-13: *LBD* distributions and *lbd* simplex for the Karate graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

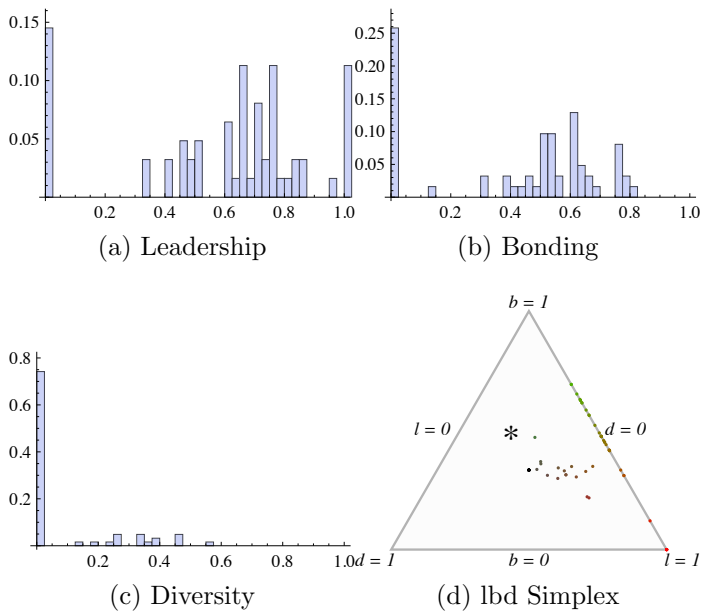
Figure A-14: *LBD* distributions and *lbd* simplex for the Karate graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

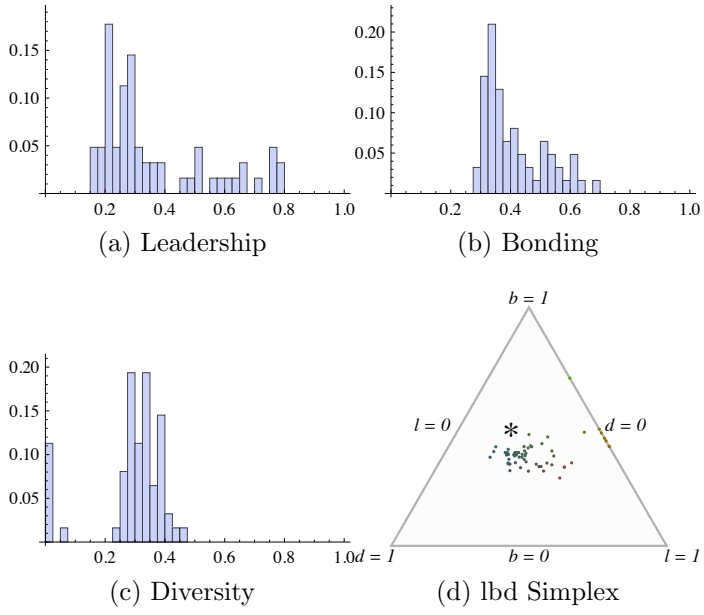
Figure A-15: LBD distributions and *lbd* simplex for the Karate graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

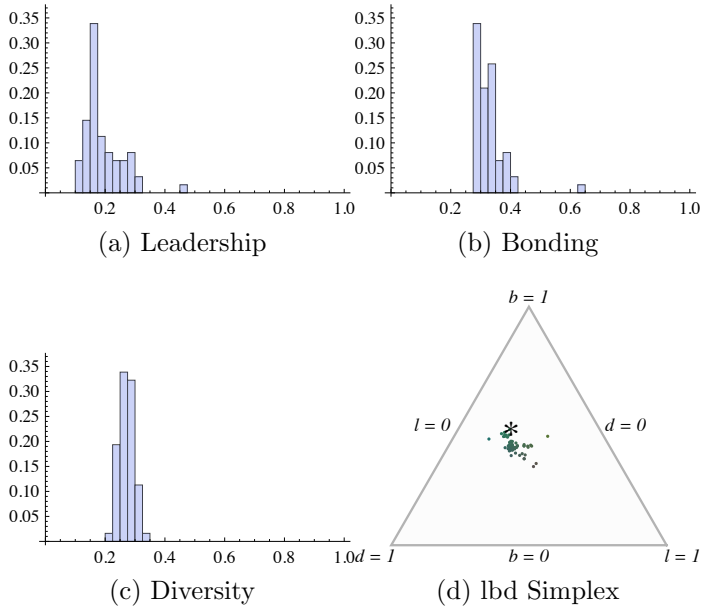
Figure A-16: *LBD* distributions and *lbd* simplex for the Dolphins graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

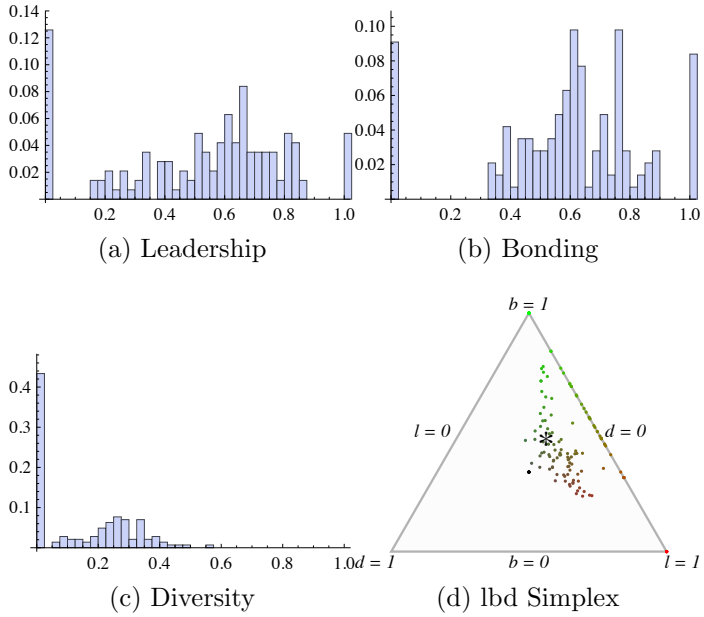
Figure A-17: *LBD* distributions and *lbd* simplex for the Dolphins graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

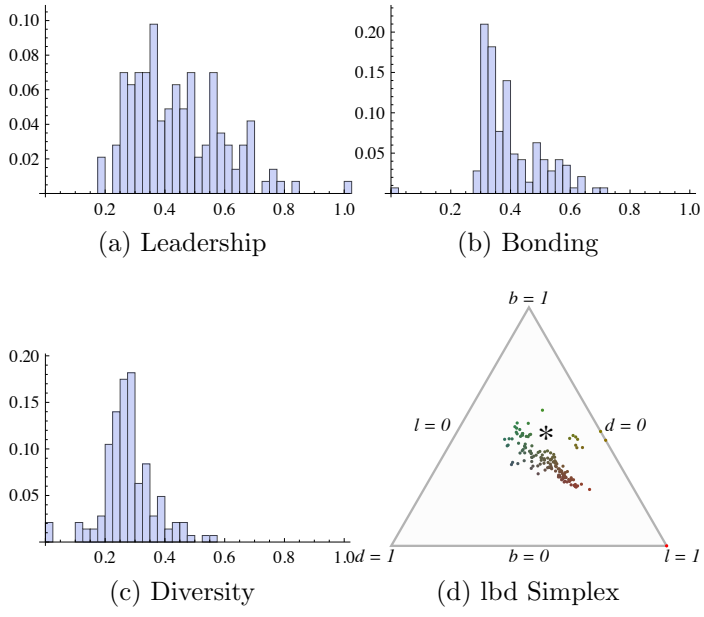
Figure A-18: LBD distributions and *lbd* simplex for the Dolphins graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

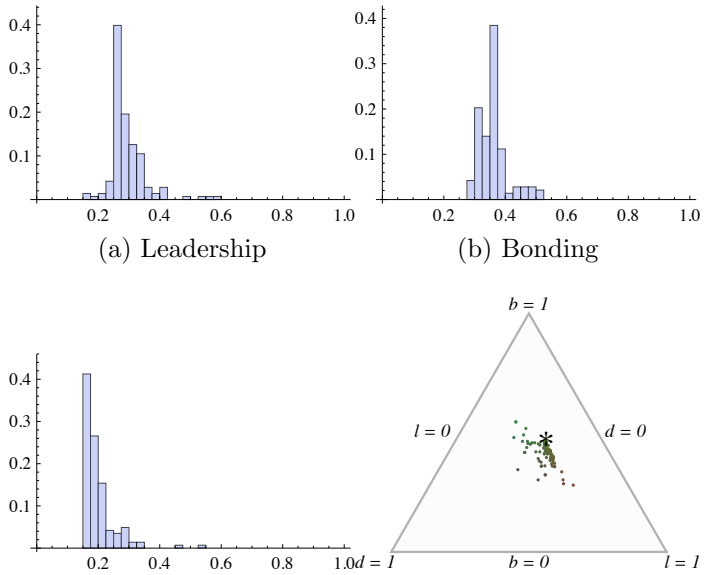
Figure A-19: LBD distributions and lbd simplex for the Enron graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

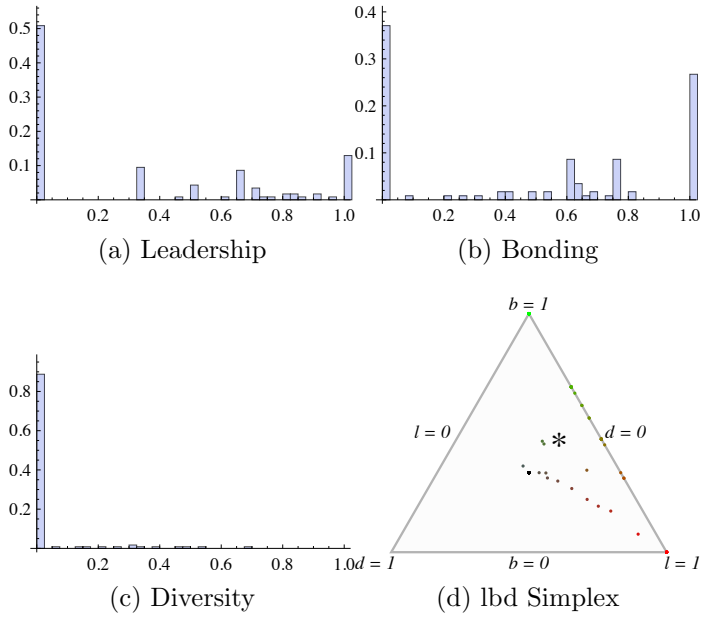
Figure A-20: LBD distributions and lbd simplex for the Enron graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

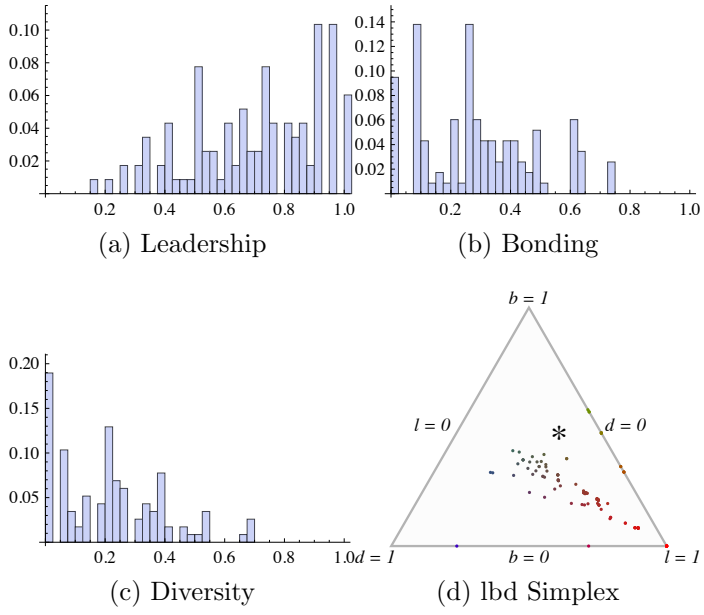
Figure A-21: LBD distributions and lbd simplex for the Enron graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

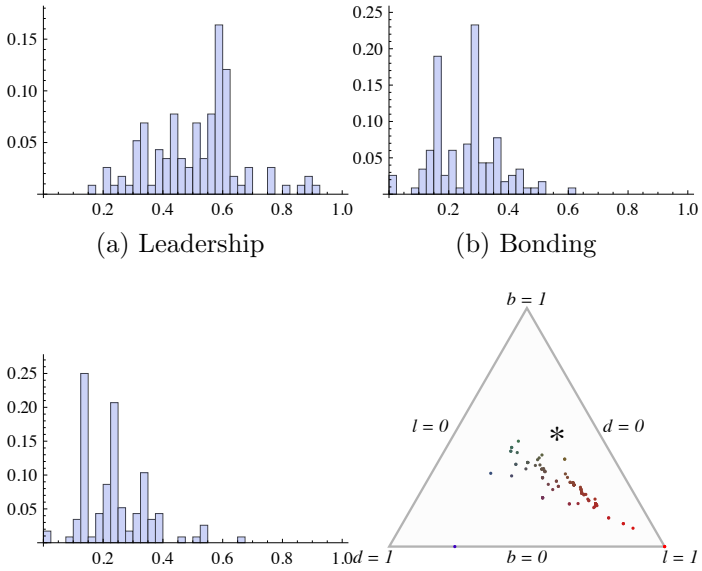
Figure A-22: *LBD* distributions and *lbd* simplex for the Santa Fe graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

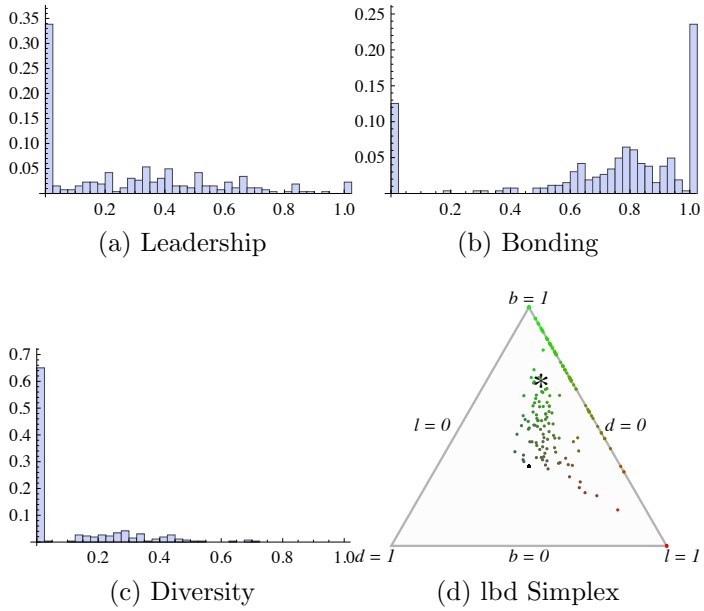
Figure A-23: *LBD* distributions and *lbd* simplex for the Santa Fe graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

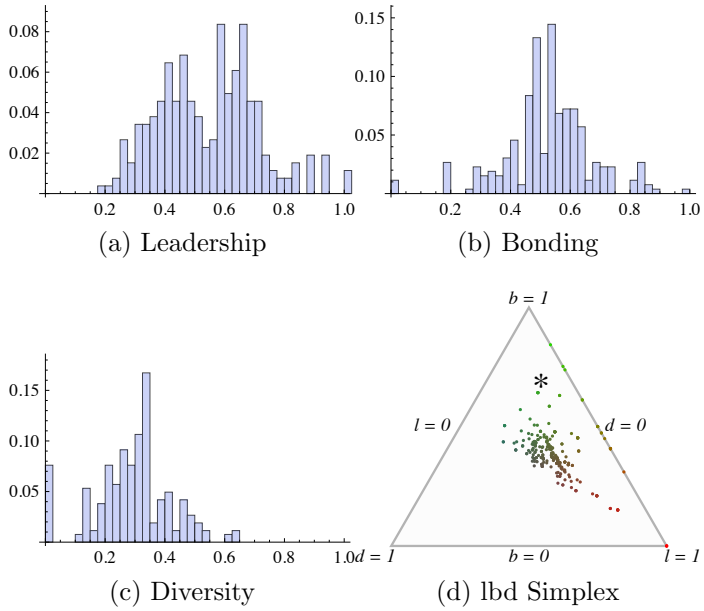
Figure A-24: LBD distributions and *lbd* simplex for the Santa Fe graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

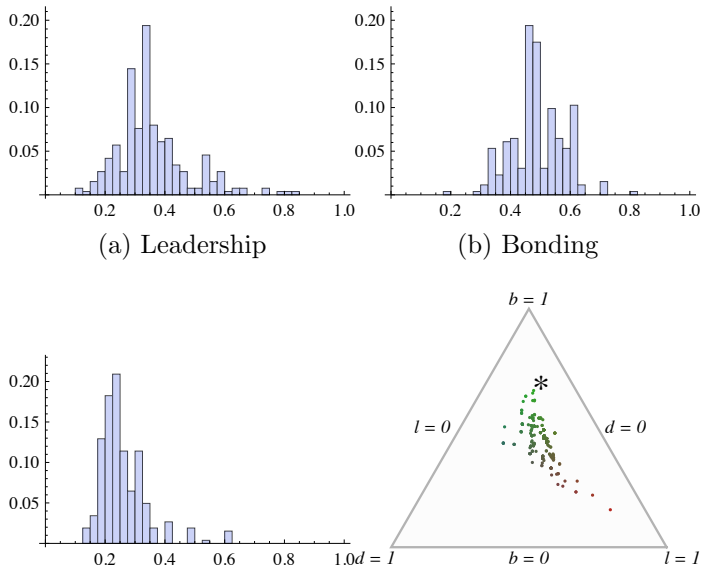
Figure A-25: *LBD* distributions and *lbd* simplex for the JJATT graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

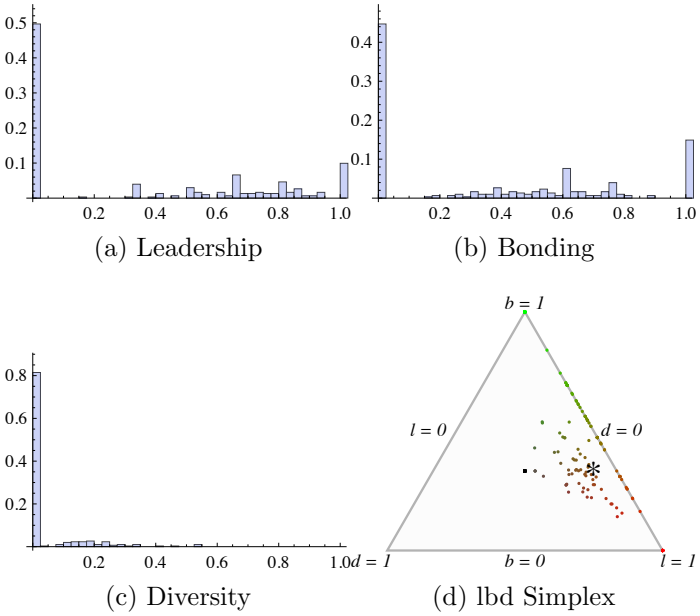
Figure A-26: *LBD* distributions and *lbd* simplex for the JJATT graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

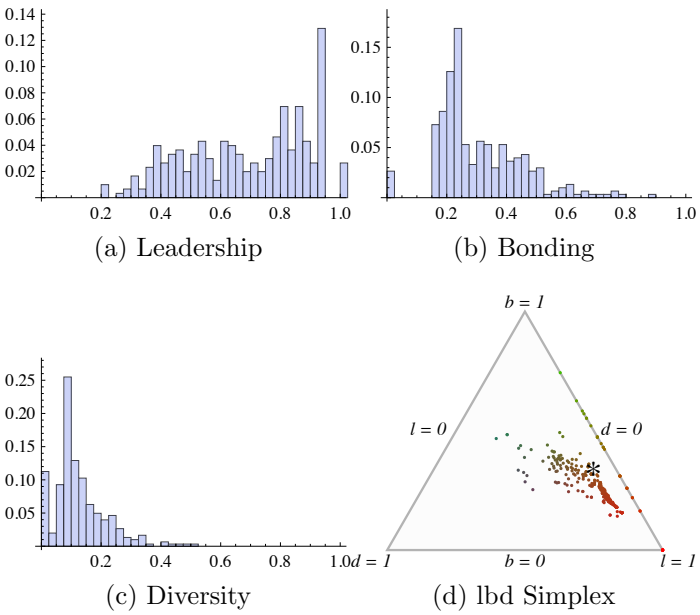
Figure A-27: *LBD* distributions and *lbd* simplex for the JJATT graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

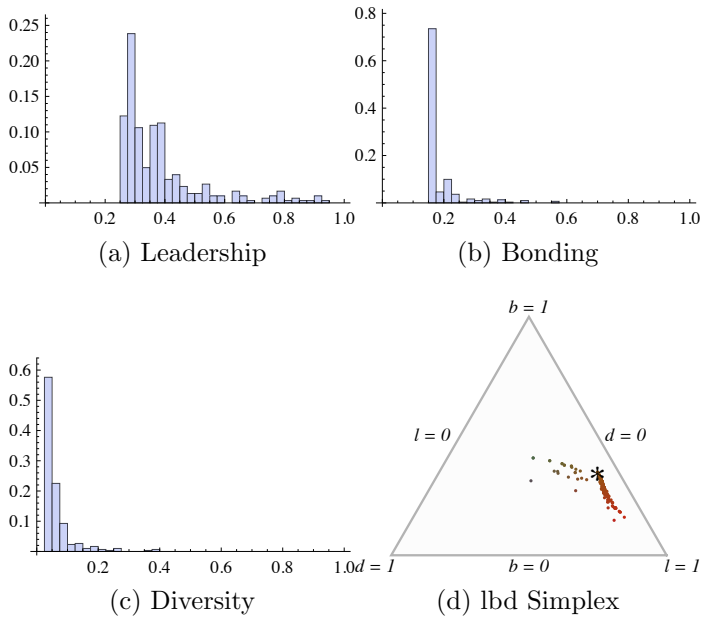
Figure A-28: *LBD* distributions and *lbd* simplex for the Linux 2001 graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

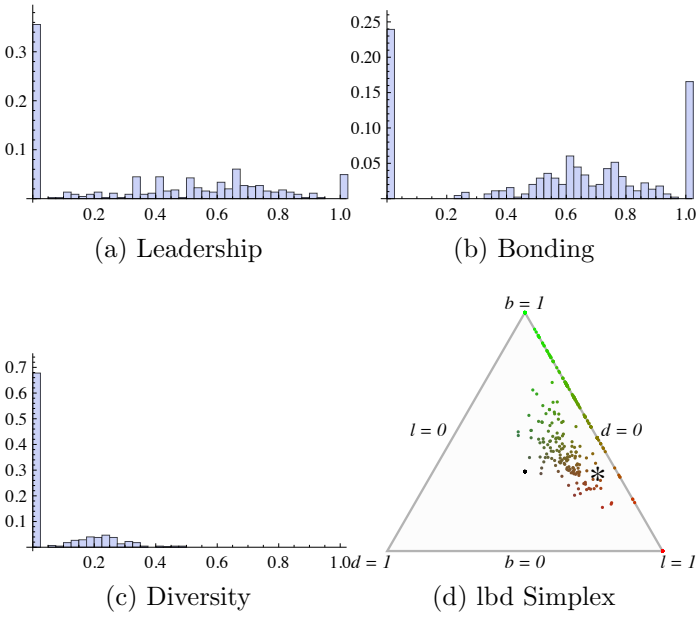
Figure A-29: *LBD* distributions and *lbd* simplex for the Linux 2001 graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

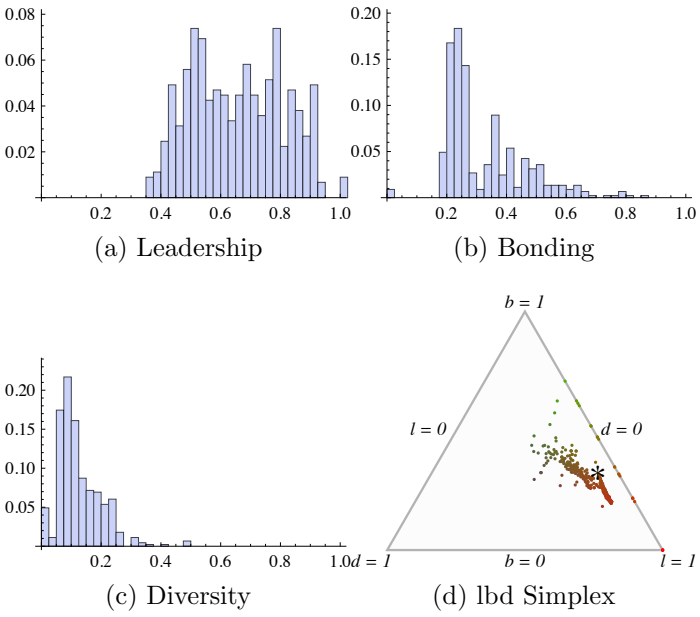
Figure A-30: LBD distributions and *lbd* simplex for the Linux 2001 graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

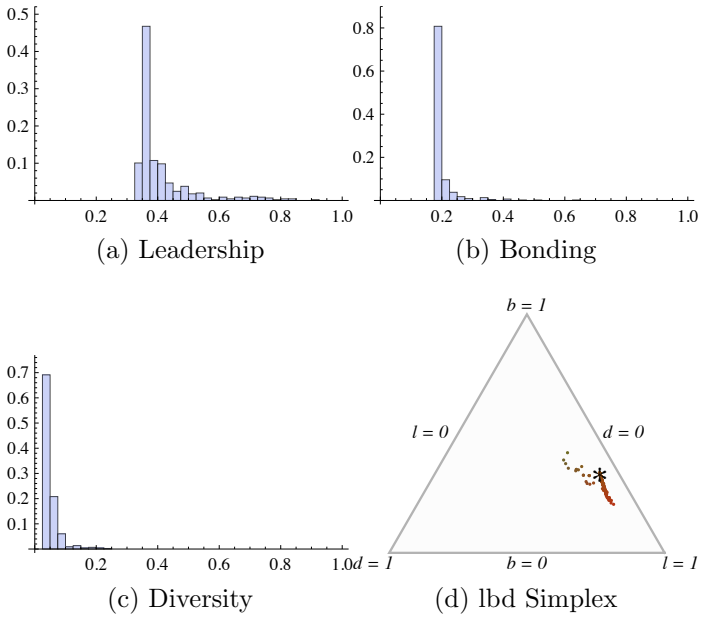
Figure A-31: *LBD* distributions and *lbd* simplex for the Linux 2008 graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

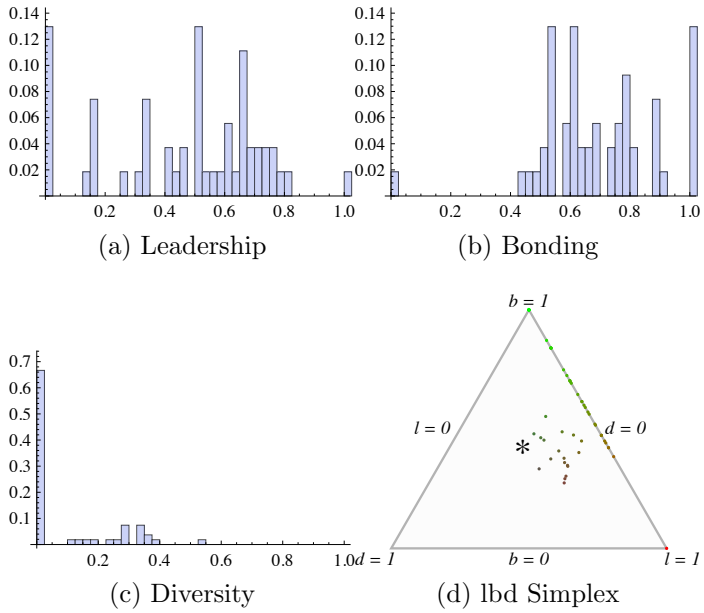
Figure A-32: *LBD* distributions and *lbd* simplex for the Linux 2008 graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

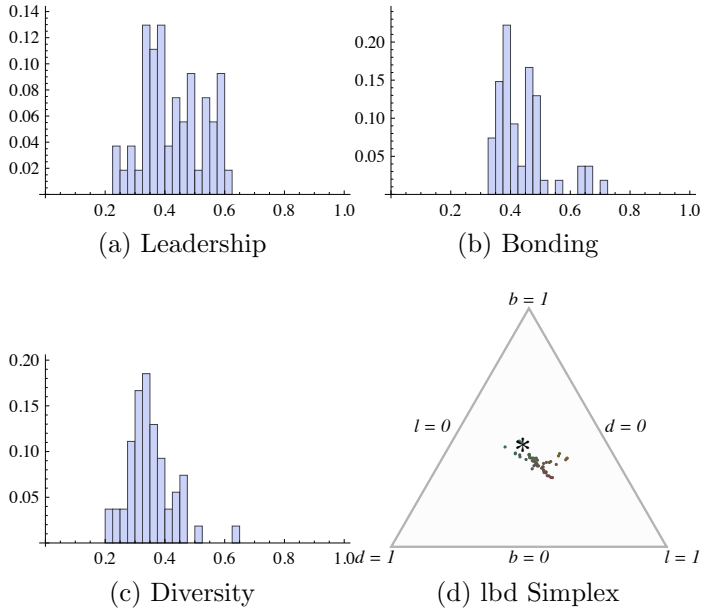
Figure A-33: LBD distributions and *lbd* simplex for the Linux 2008 graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

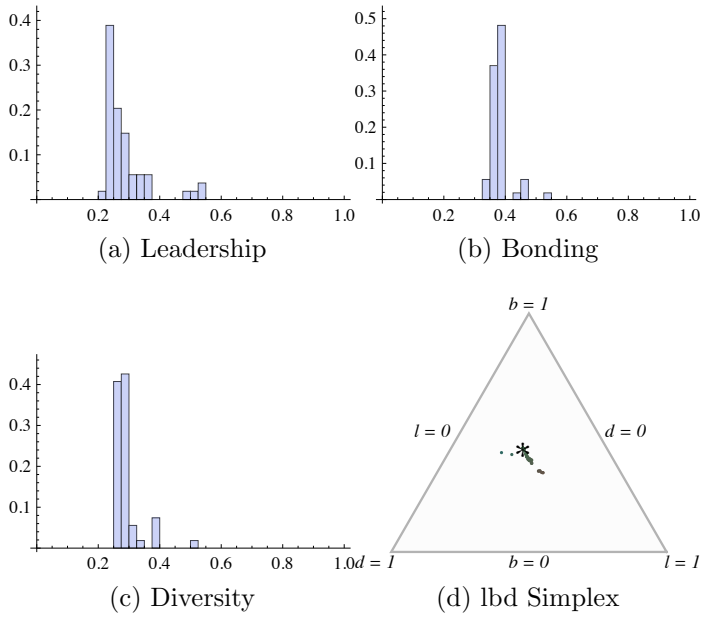
Figure A-34: LBD distributions and lbd simplex for the Bright graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

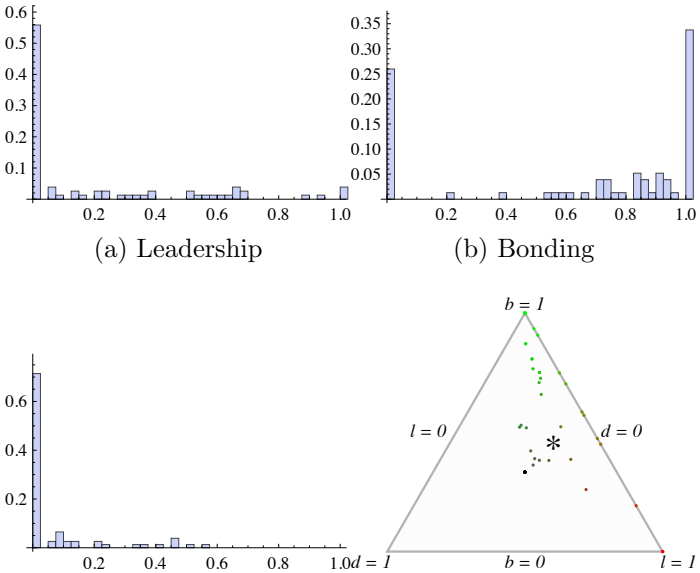
Figure A-35: LBD distributions and lbd simplex for the Bright graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

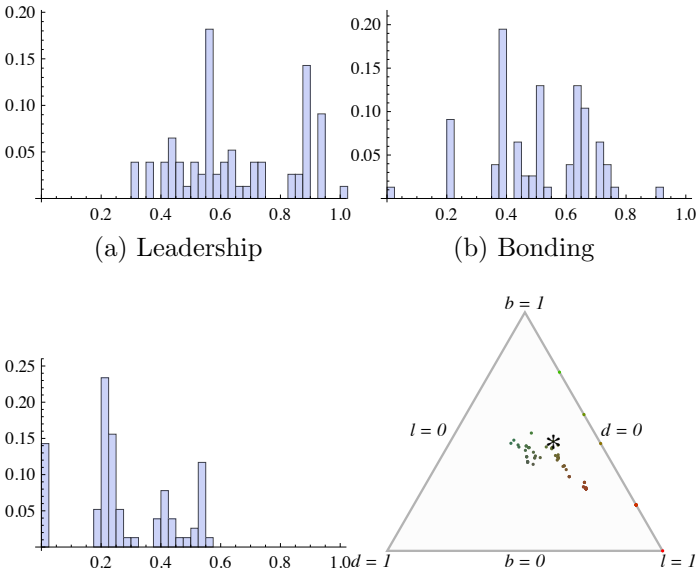
Figure A-36: LBD distributions and *lbd* simplex for the Bright graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

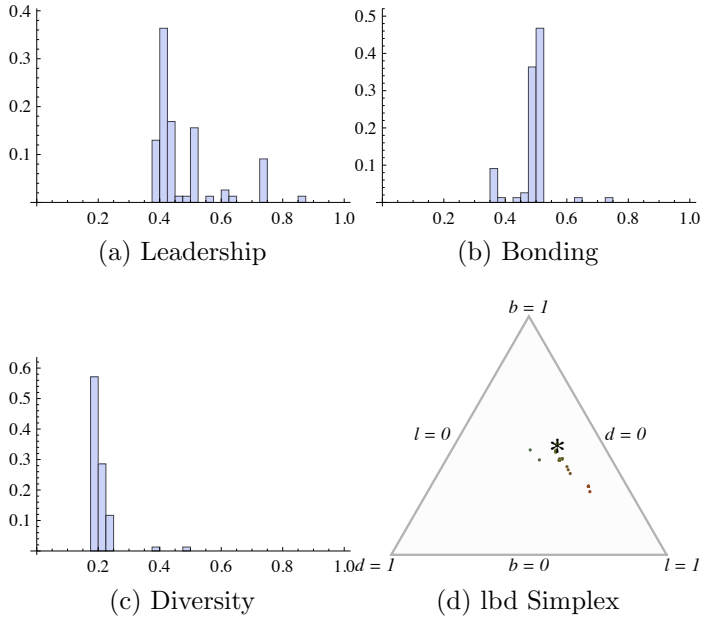
Figure A-37: *LBD* distributions and *lbd* simplex for the Lesmis graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

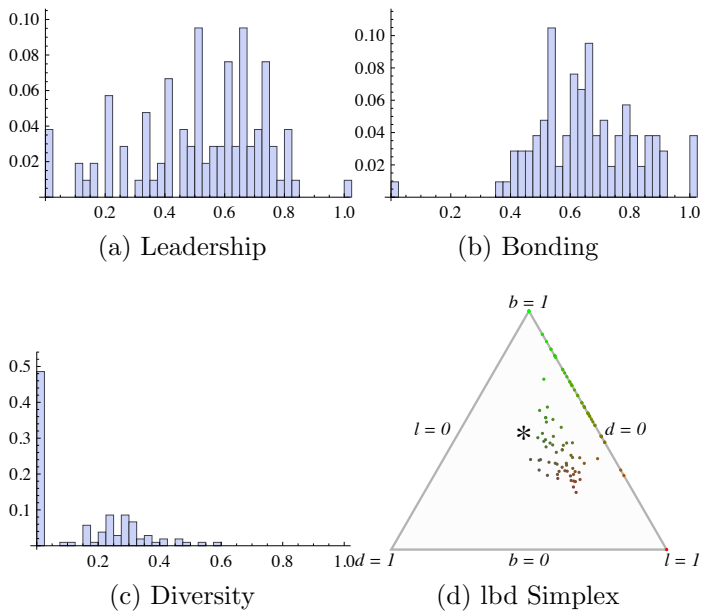
Figure A-38: *LBD* distributions and *lbd* simplex for the Lesmis graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

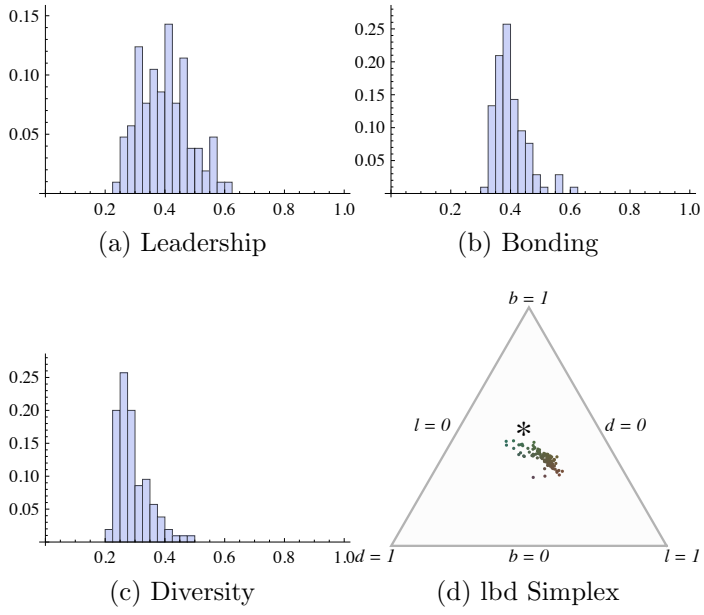
Figure A-39: LBD distributions and *lbd* simplex for the Lesmis graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

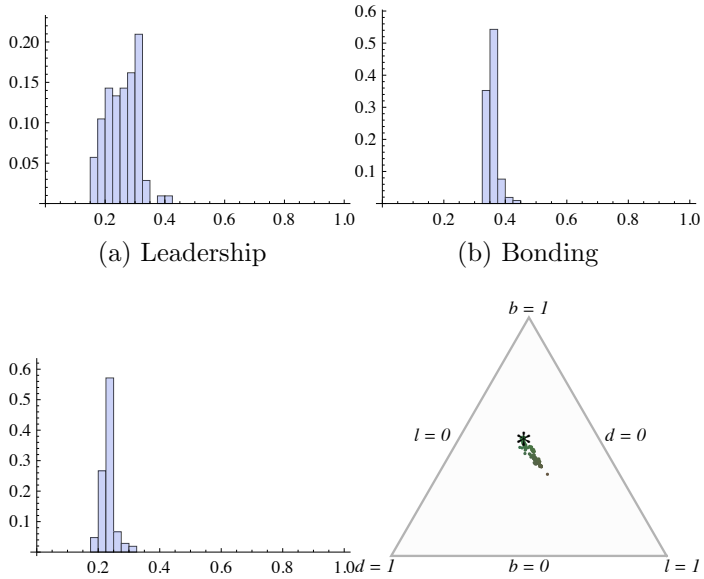
Figure A-40: LBD distributions and *lbd* simplex for the PolBooks graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

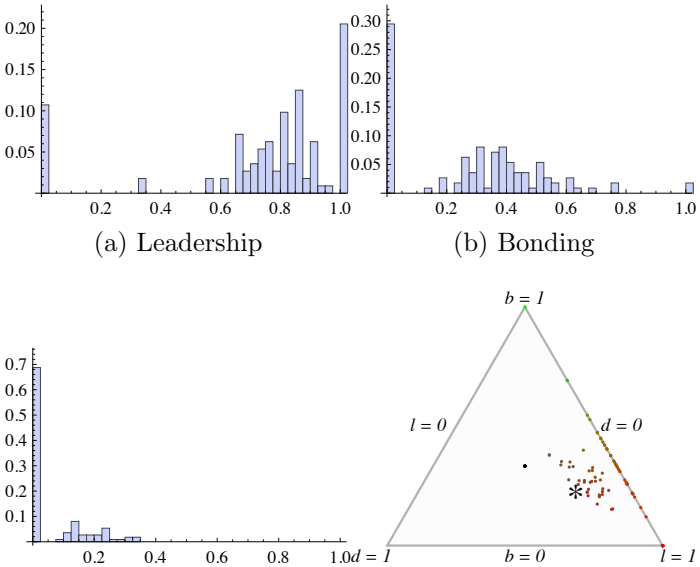
Figure A-41: LBD distributions and *lbd* simplex for the PolBooks graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

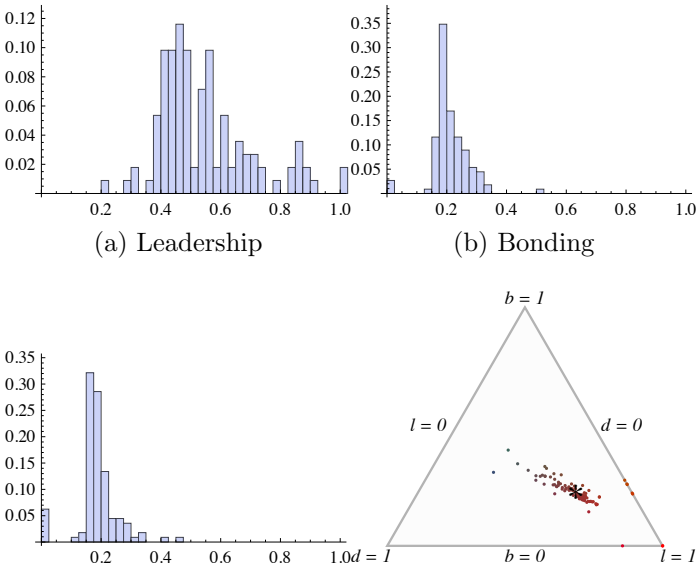
Figure A-42: LBD distributions and *lbd* simplex for the PolBooks graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

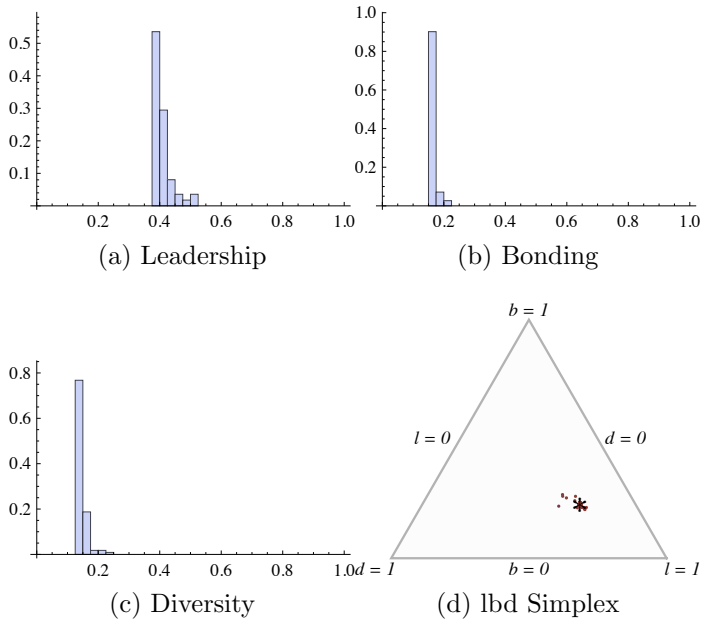
Figure A-43: *LBD* distributions and *lbd* simplex for the AdjNoun graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

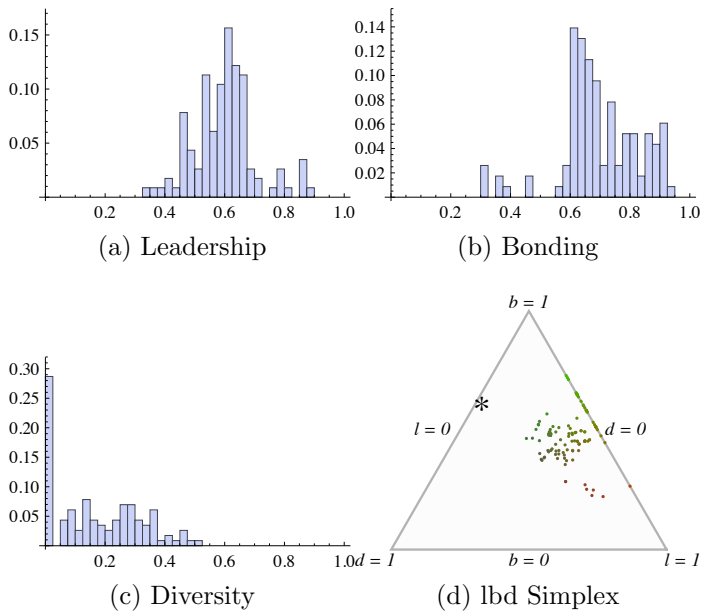
Figure A-44: *LBD* distributions and *lbd* simplex for the AdjNoun graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

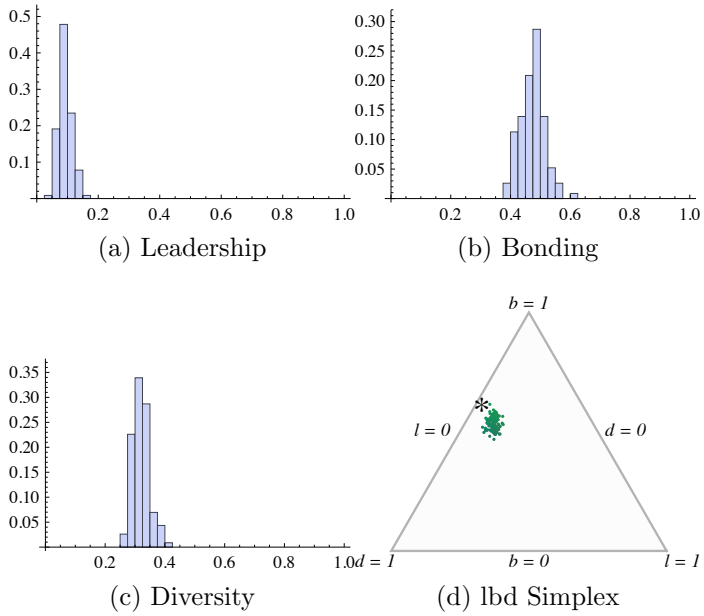
Figure A-45: LBD distributions and *lbd* simplex for the AdjNoun graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

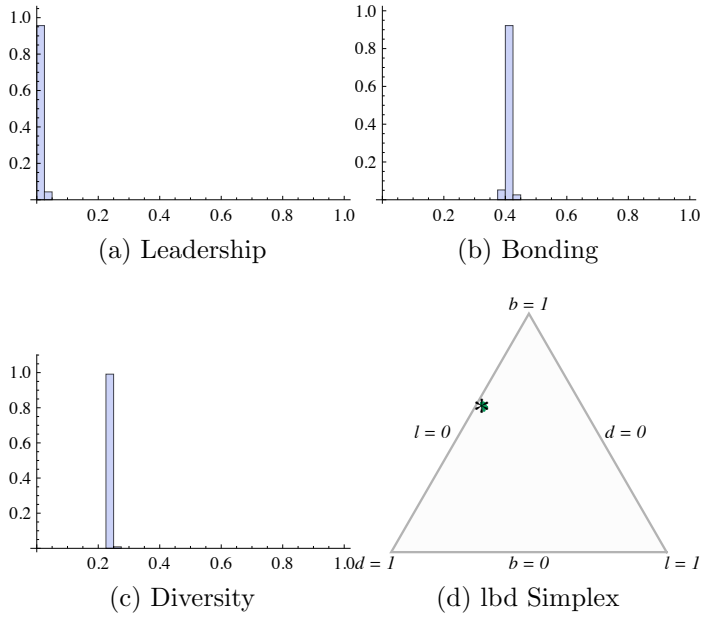
Figure A-46: LBD distributions and lbd simplex for the Football graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

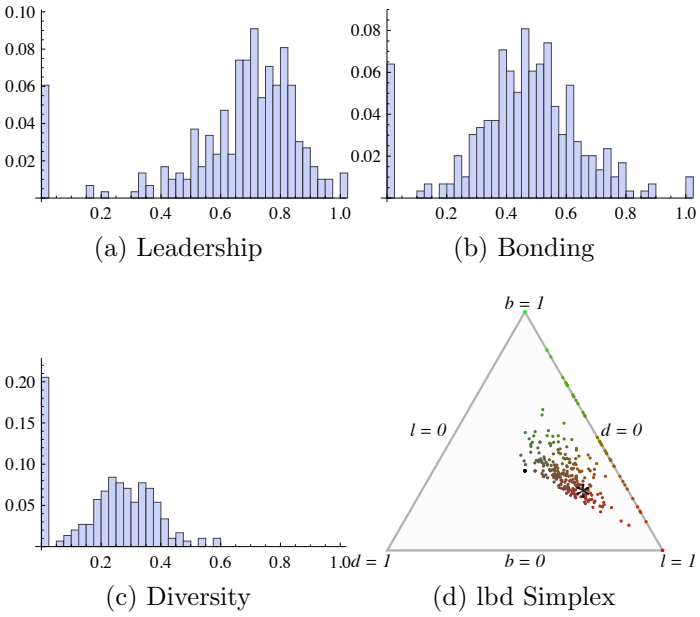
Figure A-47: LBD distributions and lbd simplex for the Football graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

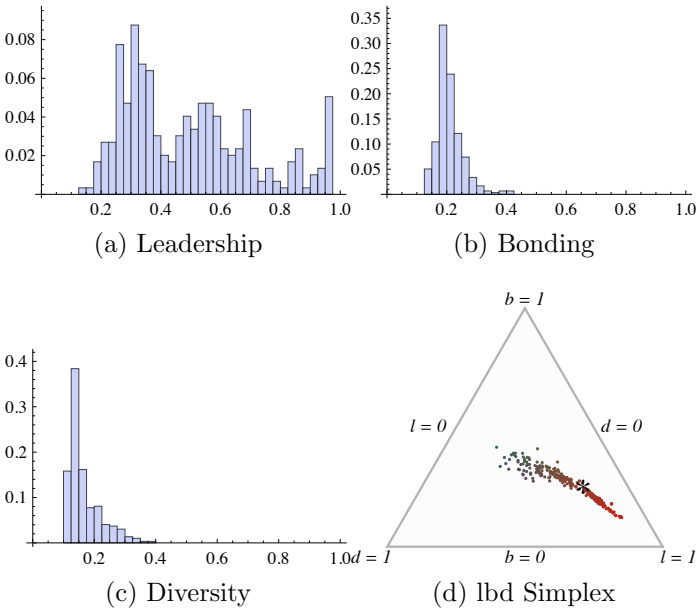
Figure A-48: LBD distributions and *lbd* simplex for the Football graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

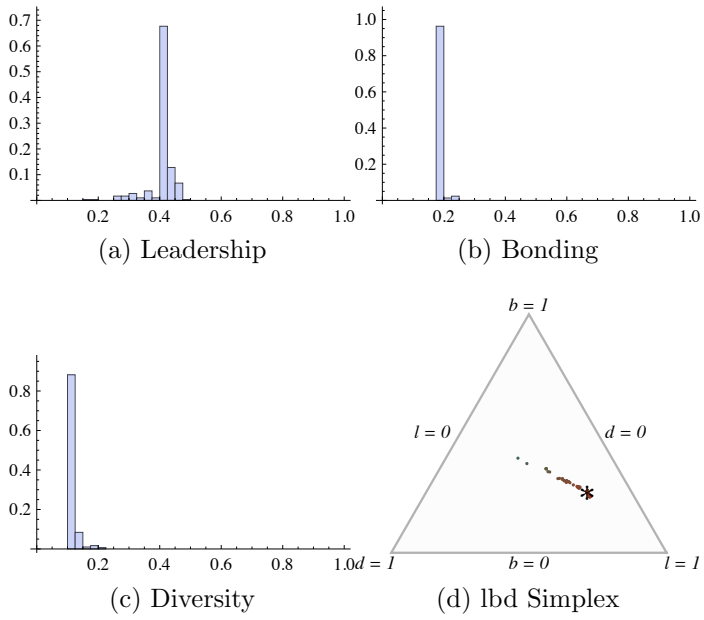
Figure A-49: *LBD* distributions and *lbd* simplex for the C-Elegans graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

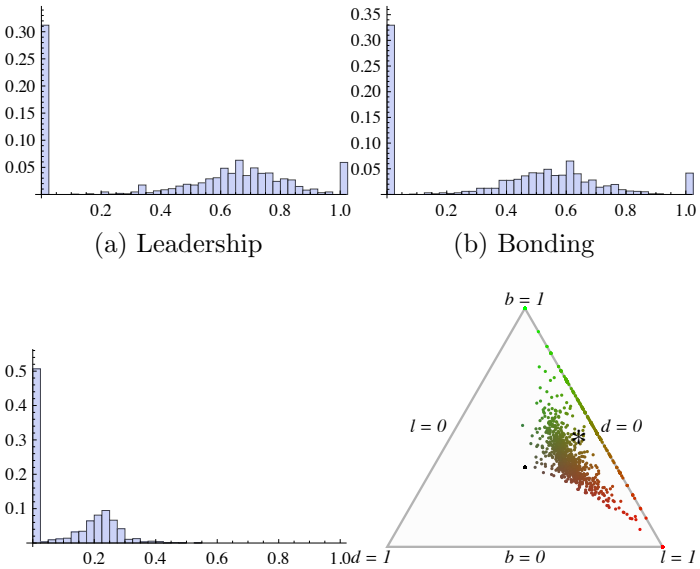
Figure A-50: *LBD* distributions and *lbd* simplex for the C-Elegans graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

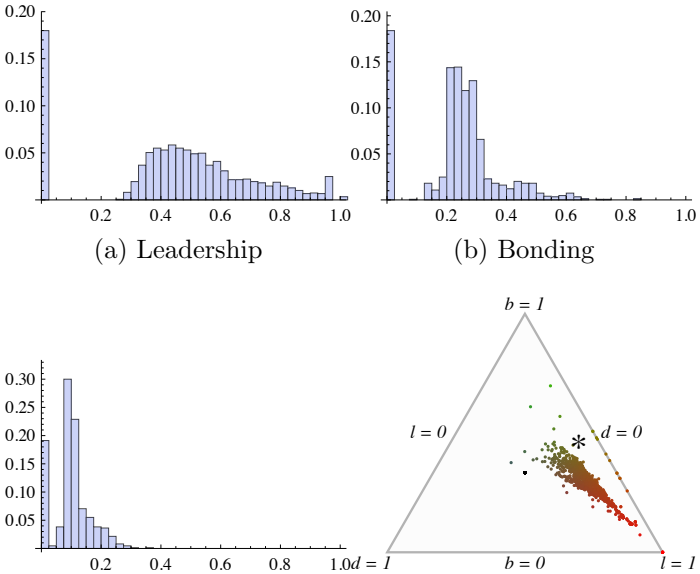
Figure A-51: LBD distributions and lbd simplex for the C-Elegans graph at radius 3.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

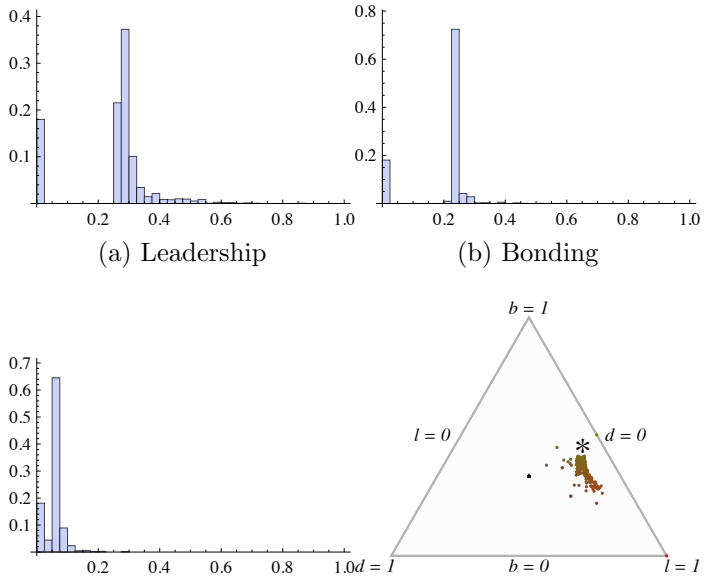
Figure A-52: *LBD* distributions and *lbd* simplex for the PolBlogs graph at radius 1.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

Figure A-53: *LBD* distributions and *lbd* simplex for the PolBlogs graph at radius 2.



(a) Leadership (b) Bonding

(c) Diversity (d) lbd Simplex

Figure A-54: LBD distributions and *lbd* simplex for the PolBlogs graph at radius 3.

Bibliography

[1] 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.

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

[3] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[4] A.-L. Barabási and R. Albert. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97, 2002.

[5] K. M. Borgwardt. *Graph Kernels*. PhD thesis, LudwigMaximilians University, Munich, 2007.

[6] G. Chartrand, F. Saba, and H. B. Zou. Edge rotation and distance between graphs. *Mathematica Bohemica*, 110:87–91, 1985.

[7] W. W. Cohen. Enron email dataset. `http://www.cs.cmu.edu/~enron/`, 2009.

[8] S.N. Dorogovtsev and J. F. F. Mendes. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford, 2003.

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

[10] P. Erd os and A. Renyi. On random graphs. *Publicationes Mathematicae*, 6:17–61, 1961.

[11] B. Everitt. *Cluster Analysis*. John Wiley, New York, 1974.

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

[13] A. Frick, A. Ludwig, and H. Mehldau. A fast adaptive layout algorithm for undirected graphs. In *Proceedings of the DIMACS International Workshop on Graph Drawing*, 1994.

[14] T. G artner, P. Flach, and S. Wrobel. On graph kernels: Hardness results and efficent alternatives. In *Proceedings of the Annual Conference on Computational Learning Theory*, 2003.

[15] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

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

[17] F. Harary. *Graph Theory*. Addison-Wesley, Reading, 1969.

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

[19] V. Krebs. Books about us politics dataset (unpublished). `http://www.orgnet.com/`, 2003.

[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*, 54:396–405, 2003.

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

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

[23] M. E. J. Newman. A set of measures of centrality based on betweenness. *Sociometry*, 40:35–41, 1977.

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

[25] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physics Review E*, 69, 2004.

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

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

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

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

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

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

[32] W. Richards and N. Wormald. Representing small group evolution. In *Proceedings of the IEEE Conference on Social Computing*, page 232, 2009.

[33] 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.

[34] B. Sch olkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

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

[36] S. Sreenivasan, R. Cohen, E. Lopéz, Z. Toroczkai, and H. E. Stanley. Communication bottlenecks in scale free networks. *Physics Review E*, 75, 2007.

[37] A. Stoica and C. Prieur. Structure of neighbourhoods in a large social network. In *Proceedings of the IEEE Conference on Social Computing*, 2009.

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

[39] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Science*, 296:440–442, 1998.

[40] 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*, 314:1–340, 1986.

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