The Use of Domain Knowledge in Optimal Information Aggregation

by

Galen Pickard

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

Master of Science in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

[June 2006]
~~May 2006~~

Author ....
Department of Electrical Engineering and Computer Science
May 26, 2006

Certified by ....
Whitman Richards
Professor
Thesis Supervisor

Accepted by ....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

The Use of Domain Knowledge in Optimal Information Aggregation

by

Galen Pickard

Submitted to the Department of Electrical Engineering and Computer Science
on May 26, 2006, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science and Engineering

Abstract

In this thesis, I present some novel results pertaining to the relationship between two popular and interesting information aggregation methods: the Condorcet and Borda tallies. I present numerical results showing how the much simpler Borda tally can be used to approximate the outcome of the Condorcet tally with high probability in certain circumstances, a proof that there exist classes of problems for which the two tallies can never agree, and an extension of these results to small-world graphs, which have been of great interest recently due to their practical applicability to many complex problems.

Thesis Supervisor: Whitman Richards
Title: Professor

Contents

List of Figures

List of Tables

Chapter 1

Introduction

At its most fundamental, an information aggregation problem is one in which information being presented from multiple sources must be combined into a single output. These problems appear in a wide variety of domains. In distributed systems, one master device may query a large number of other devices, and then combine their responses into a single output. In algorithms, multiple heuristics for finding the solution to a problem may find different results, and some strategy is needed to choose between the various alternatives. In the social sciences, leaders rely on cabinets of advisors, whose individual inputs must be coalesced into a single policy decision. Cognitive processes can be described as the result of the combined outputs of many internal neuronal "experts," which evaluate various facets of the world[6]. Perhaps most directly, in actual practice, whenever members of a democratic organization vote, the voting system is an information aggregator which takes the opinions of the voting populace and merges them into a choice among some set of candidate alternatives[8].

Chapter 2

Condorcet and Borda Tallies

2.1 Definitions

We are primarily interested in comparing the behaviors of two information aggregation methods, namely the Condorcet[3] and Borda[2] tallies. This interest is motivated by both historical and practical concerns: the relative merits of the two methods have been argued since the 18th century, and, although variations on the Borda tally are much more common in practice, there is a significant body of theoretical evidence that the Condorcet tally is in many cases superior[10]. To begin, we provide formal definitions of both of these methods, and of a generalized information aggregation problem.

An information aggregation problem, often referred to simply as a voting problem, is a problem which involves taking information from a number of disparate sources, and using this information to produce a single output. Problems of this type arise in numerous domains. In political science, they are common, occurring whenever voters cast votes which are used to decide the outcome of some issue. Although less directly, they also occur in fields as diverse as control theory and cognitive science. Wherever there is a system which must coalesce information from its subsystems, there is potentially the need to balance conflicting information about a single topic. In any such circumstance, if there is not reason to trust any single source more than the others, the problem can be phrased as one of information aggregation.

2.1.1 Information Aggregation Problems

To facilitate the formalization of these problems, we only consider cases in which a finite number of information sources (which shall henceforth be referred to as voters) provide information regarding some contested issue. Furthermore, we only consider cases in which there is a finite set of candidates which the voters are evaluating and choosing among. It is to be understood that these candidates should not necessarily be interpreted in the political sense of persons competing for an office, but rather, in a more general sense, as potential solutions for whatever the issue at hand is (for example, in a control system, the candidates might be the set of allowed actions, i.e. $\{forward, reverse, right, left\}$). We must also formalize the manner in which voters provide information about the relative fitness of the candidates. Although other options exist, we will limit a voter to providing a partially ordered list of the candidates. The output of the information aggregation method, then, is a partially ordered list which is meant to represent the overall beliefs of the voting populace as a whole. A partially ordered list can be described using the relationships $\succ$ and $\sim$, which should be read as "is preferred to" and "is considered equal to," respectively. The $\succ$ relationship is transitive, and the $\sim$ relationship is reflexive, symmetric, and transitive. Also, $A \succ B, B \sim C \Rightarrow A \succ C$. One can think of such a partially ordered list as a way of dividing candidates into "tiers," such that every member of a given tier is preferred to every member of any lower ranked tier, and such that all member of a given tier are considered equal. For example, consider the following partially ordered list of candidates $\{A, B, C, D, E, F\}$: $A \succ B \sim C \succ D \sim E \succ F$. This list would represent the opinion of a voter who considered A to be the best candidate, B and C to be tied as equally fit candidates which were both worse than A, but which were preferred to all others, D and E to be similarly tied candidates which were superior only to F, and F to be the least desirable candidate.

In many cases, especially in political science, voters vote for a single candidate, rather than providing an ordering. This, however, represents a specialized ordering, in which a voter who casts a vote for candidate X can be seen as simply providing the ordering $X \succ A \sim B \sim \ldots$. Similarly, if the outcome of the information aggregation is a single choice, rather than an ordering, any ordering with a single most preferable choice can be truncated to a similar format. Thus, this representation is sufficiently powerful to encompass a wide range of problems with relative ease.

2.1.2 Condorcet Tally

Now, using this formal framework for describing information aggregation problems, we will describe two possible solutions: the Condorcet and Borda tallies. The Condorcet and Borda tallies are two useful and historically significant aggregation methods, both of which date back to the 18th century. The Condorcet tally is designed to satisfy the Condorcet criterion: if the majority of voters prefer candidate A to candidate B, then $A \succ B$ in the outcome of the tally. The tally follows naturally from the criterion - for each pair of candidates, all voters are queried as to which is preferred. The candidate which is preferred by more voters is the winner of that pairwise comparison. The overall ordering is uniquely determined by the set of outcomes of all such pairwise comparisons. There is a fundamental problem with this method, however, insofar as that it is possible for candidate A to beat candidate B in a pairwise comparison, candidate B to beat candidate C, and yet, at the same time, candidate C beat candidate A. In a case such as this, clearly, no transitive ordering can be produced which satisfies all of these pairwise constraints.

To make this concrete, consider a problem involving trying to find an ordering of three candidates: $\{rock, paper, scissors\}$. Three voters are queried, and they provide the following three orderings: $rock \succ scissors \succ paper$, $scissors \succ paper \succ rock$, and $paper \succ rock \succ scissors$. To calculate the outcome of the Condorcet tally, we must calculate the pairwise outcomes between *rock* v. *paper*, *rock* v. *scissors*, and *paper* v. *scissors*. When we do this, we find that $\frac{2}{3}$ of the voters prefer *paper* to *rock*, so we must have $paper \succ rock$ in the overall ordering. Similarly, $\frac{2}{3}$ of the voters prefer *rock* to *scissors*, and $\frac{2}{3}$ prefer *scissors* to *paper*, so $rock \succ scissors$ and $scissors \succ paper$. However, all three of these cannot be simultaneously satisfied, as we require $\ldots \succ rock \succ scissors \succ paper \succ rock \succ \ldots$. Thus, there is no ordering which satisfies the Condorcet criterion, given these candidates and voters' preferences.

2.1.3 Borda Tally

In a Borda tally, each voter provides a ranked ordering of all candidates, in the form $A \succ B \succ C \succ \ldots$. Then, for each voter, each candidate is assigned some number of points. Classically, the number of points assigned follows a linear progression from $n-1$ to 0, where n is the number of candidates. However, in general, the number of points assigned could be any vector B of length n, subject to the constraints that $\forall i . B_i \geq B_{i-1}$ and $B_0 > B_{n-1}$. We call this vector a Borda vector. If we wish to allow a voter to submit a partial ordering of the candidates, in which two candidates are allowed to tie, we need to modify the tally slightly. There are three obvious modifications one could make to accomplish this. Supposing we have a Borda vector $B = \{B_0, B_1, B_2, \ldots, B_{n-1}\}$, and a voter whose partial order is $A \succ B \sim C \succ D \succ \ldots$. In either case, A receives B_0 points. One option for handling the tie between B and C is to assign them both $\frac{B_1+B_2}{2}$ points. D, then, would receive B_3 points. Another option is to assign both B and C B_1 points. Then, D could be assigned either B_2 or B_3 points. Among these 3 variations, the first is truest to the spirit of the Borda tally, as each voter gives out the same total number of points. However, as will become apparent later in this thesis, we have good reason to consider using the other modifications in certain specialized domains.

2.2 Arrow's Theorem

One might ask, at this point in the discussion of information aggregation, whether there is a single aggregation method which is universally preferable to all others. Unfortunately, there is not - it is a classic result known as Arrow's Theorem that, given a (relatively simple) set of desirable traits for an information aggregation method, there can be no method which simultaneously satisfies all of them[1]. Thus, there are necessarily situations in which any given method will act in an undesirable manner. There are multiple such sets of properties, but a particularly illustrative one is as follows:

1. Universality: The aggregation method should be produce a valid partial ordering as its outcome, should be deterministic, and every voter should be allowed to provide any valid partial ordering as a preference order.

2. Non-imposition: Every valid ordering should be achievable as an outcome via some set of voter inputs.

3. Non-dictatorship: There should not be a "special" voter whose ordering uniquely determines the outcome of the aggregation method.

4. Pareto efficiency: If every voter prefers candidate A to candidate B, then $A \succ B$ in the outcome of the aggregation method. (Note that this criterion is simply a stronger version of the Condorcet criterion.)

5. Independence of irrelevant alternatives: Given a set of candidates and voters, the introduction of a new candidate should not change the ordering of the original candidates relative to each other.

To give the reader a more intuitive understanding of the last trait, we note that it is violated by the "plurality" method used in the American political system, in which each voter casts a single vote, presumably for their favorite candidate. This violation was made famous in the 2000 presidential elections, in which the availability of Ralph Nader as a candidate was seen by many to have altered the relative ranking of George Bush and Al Gore. Regardless of one's feelings on the end result of this particular election, it should be apparent that this particular feature was a flaw in this case, as it led to widespread situations in which voters whose true preference order was $Nader \succ Gore \succ Bush$ ended up casting their votes as though Gore were their first choice. This is referred to as "voting strategically," and should be taken as a fault - voters should be able to vote according to their true beliefs, and trust the aggregation method to take them into account appropriately.

We note, then, that this plurality method is simply a special case of a Borda tally, in which the Borda vector used is $B = \{1, 0, 0, \ldots\}$. The dependence on irrelevant alternatives is not limited to a plurality tally, and is, in fact, a property of any Borda tally. It should be readily apparent, however, that the Borda tally satisfies the other four properties listed above. By contrast, the Condorcet tally does satisfy the independence of irrelevant alternatives property. Since candidates are only compared in a pairwise fashion, the presence of other candidates cannot influence the relative rankings of any two given candidates. The downfall of the Condorcet method is that it violates universality, as described earlier - in the $\{rock, paper, scissors\}$ example, the aggregation method fails to arrive at a valid result.

2.3 Graphical Knowledge Representations

For the majority of this paper, we will be discussing a specific subset of information aggregation problems: those for which voters adhere to an underlying shared knowledge structure that can be represented in the form of a graph[5]. We consider a graph to be a collection of nodes connected by edges. We allow nodes to be weighted, such that every node has an associated weight w, and we consider cases in which edges are unweighted and bi-directional. More complicated knowledge structures can be represented by allowing weighted and/or directed edges, but those problems are beyond the scope of this thesis. In the problems addressed here, the nodes generally represent both the candidates and the voters. Each node has associated with it a partial ordering, and the weight of that node is simply the number of voters who supply that particular partial ordering. It follows, then, that for a graph with n nodes, there are at most n unique partial orderings available to voters. Thus, the restriction imposed by a graphical model inherently violates non-imposition for any $n > 2$, since there are $n!$ valid orderings, but only n can be supplied by voters.

The orderings associated with each node are a function of the graph structure. In these representations, edges are used to signify some sort of similarity relationship between two nodes[4]. The exact interpretation of this notion varies depending on the domain, but general meaning is straightforward. Thus, a node N is more similar to its neighbors than it is to nodes which are not connected to it. By extension, then, a node N is more similar to neighbors of its neighbors than it is to any nodes which are not neighbors of N, or of any of its neighbors. We speak, then, about the distance between two nodes as the length of the shortest path connecting them, with the distance between any node and itself being 0. It follows that any node of distance k from N is mode similar to N than any node of distance $l > k$. We make the assumption, then, that the preference order for a node should follow these similarity comparisons, such that a node prefers other nodes which are more similar to itself. Thus, in a case where nodes A, B, C, D, E were distances $1, 1, 2, 2, 3$ from node N, respectively, the partial order provided by N would be $N \succ A \sim B \succ C \sim D \succ E$.

The method by which we apply the Condorcet tallies to graphs is relatively straightforward. For the Condorcet tally, for every pair of nodes, every node assigns points equal to its weight to whichever of the two is closer to itself, or assigns no points at all if it is equidistant from both. Then, the candidate node with the higher number of points is the winner of that pairwise comparison, and the rest of the tally proceeds as above. This follows directly from the preference orderings described above. Borda tallies, on the other hand, are a bit more complex. For any connected graph with $n > 2$ nodes, there will necessarily be some node whose partial order contains $\sim$ relationships. Thus, we must address the problem posed in section 2.1.3, regarding how to properly modify the Borda tally to deal with these. To reiterate, suppose we have a Borda vector $B = \{B_0, B_1, B_2, \ldots, B_{n-1}\}$, and a voter whose partial order is $A \succ B \sim C \succ D \succ \ldots$. The question at hand is the handling of the tie between B and C. The following are the three most reasonable options:

1. Assign $\frac{B_1+B_2}{2}$ points to both B and C. Assign B_3 points to D.
2. Assign B_1 points to both B and C. Assign B_2 points to D.
3. Assign B_1 points to both B and C. Assign B_3 points to D.

As mentioned earlier, option 1 is truest to the spirit of the Borda tally. However, it does not interact nicely with the graphical domain. In particular, it violates the fifth constraint from Arrow's theorem: independence of irrelevant alternatives. When evaluating how many points to assign to a node, a voter must consider not only the nodes on the path between its own and the one in question, but also all nodes which are no farther than away than the node in question. In many situations where we are using these graphs to describe similarity, this is undesirable behavior, and also increases the computational complexity of the Borda operations greatly. The same problem applies to option 3, since in order to compute the number of points to assign to D, one must count all nodes which are closer than D. Thus, we will use option 2 as our chosen modification of the Borda count as applied to graphical models. For the Borda tally with some Borda vector B, each node assigns points to each node depending only on the distance between the two: for a node N of weight w, N is assigned wB_0 points, all neighbors of N are assigned wB_1 points, all nodes of distance 2 are assigned wB_2 points, and so on. Then, the final order is simply the set of nodes, ordered by total number of points received.

These types of models lend themselves naturally to a number of applications. If the graph is a social network and all weights are set at 1, it is a good representation for a group selecting a leader from amongst themselves. If the candidates are more abstract choices, such as in a cognitive system or a control system, then using edges as a "similarity measure" and node weights as input weights, graphical knowledge structures represent domain information which can be used in the information aggregation process. We note that only a subset of all possible problems can be phrased in terms of a graphical knowledge structure, so it is a constraint on various aspects of the aggregation problem. This constraint will allow us to produce some interesting results which are not necessarily true for problems which are not based on a graphical knowledge structure.

Chapter 3

Prediction Between Borda and Condorcet

Here, we address the question of whether the outcome of a Condorcet tally can be approximated using methods based on Borda tallies. The motivation for this is two-fold: first, the outcome of a Condorcet tally is generally expensive to compute directly, so being able to use other methods to get the outcome would be desirable. Second, a Borda tally will always produce a valid transitive orderings, but a Condorcet tally is capable of producing orderings which violate transitivity (i.e. $A \succ B \succ C \succ A$). Clearly, then, there are outcomes of a Condorcet tally which Borda tallies are incapable of replicating. In cases where the outcome of a Condorcet tally is transitive, however, that ordering is optimal according to a number of relatively strict standards. Thus, a method based on Borda tallies which agreed with a Condorcet tally in those cases in which the Condorcet tally would have produced a transitive output would be a powerful tool.

3.1 Numerical Approximation

To begin our exploration of the relationship between the Borda and Condorcet tallies in the domain of problems which can be described by graphical representations, we present some novel numerical results, based on Monte Carlo simulations. In particular, we will examine whether there exist families of graphs for which we can use Borda tallies as a good estimator of the outcome of the Condorcet tally. We note that, although we will be discussing probabilities of agreement of the outcomes of the two tallies, they are both deterministic. When we speak of probability, we mean it with regards to a family of random graphs, with the understanding that, for any given graph, the tallies either do or do not agree.

We begin by addressing the topic of dense random graphs with a fixed edge probability $0 < p < 1$. These graphs are generated by taking a set of n nodes, then considering every pair of nodes. For every such pair, with probability p, an edge is added connecting the two nodes. These graphs are dense, since there are $O(n^2)$ edges, in expectation. As discussed earlier, it is possible that no unique Condorcet winner exists, and instead the Condorcet tally returns an invalid, non-transitive ordering. For now, we will consider only cases in which there is a Condorcet winner, and we will seek to estimate it using a Borda tally. We will examine Borda vectors of the form $B = \{1, b, 0, 0, \ldots\}$, both because it is experimentally simpler to vary only one parameter, but also because, as $n \rightarrow \inf$, the diameter of a dense random graph with fixed edge probability approaches 2 with high probability.

Generating an instance of a dense random graph, we can use linear algebra to compute the range of values of b for which the Condorcet tally and Borda tally using $B = \{1, b, 0, 0, \ldots\}$, are in agreement. Using a Monto Carlo methodology, we can estimate the probability of agreement for any given value of b, for some n and p. Figure 3-1 shows the results of such an experiment, for $n = 100, p = \{.1, .5, .9\}$. Results are similar for a wide range of n, and are elided for brevity.

These results are especially clear for the $p = .9$ case. For these graphs, using a Borda vector of $B = \{1, .5, 0, 0, \ldots\}$ correctly predicted the Condorcet winner with a failure probability of less than .01%. This Borda vector is found to be the best estimator for all tested $p \geq .4$, and at least as good as any other for all tested $p < .4$, many of which result in a large range of p between .5 and 1 which are statistically indistinguishable in terms of predictive power. This fits with the results of previous work, in which using a scheme similar to a Borda vector of $B = \{1, .5, 0, 0, \ldots\}$ was

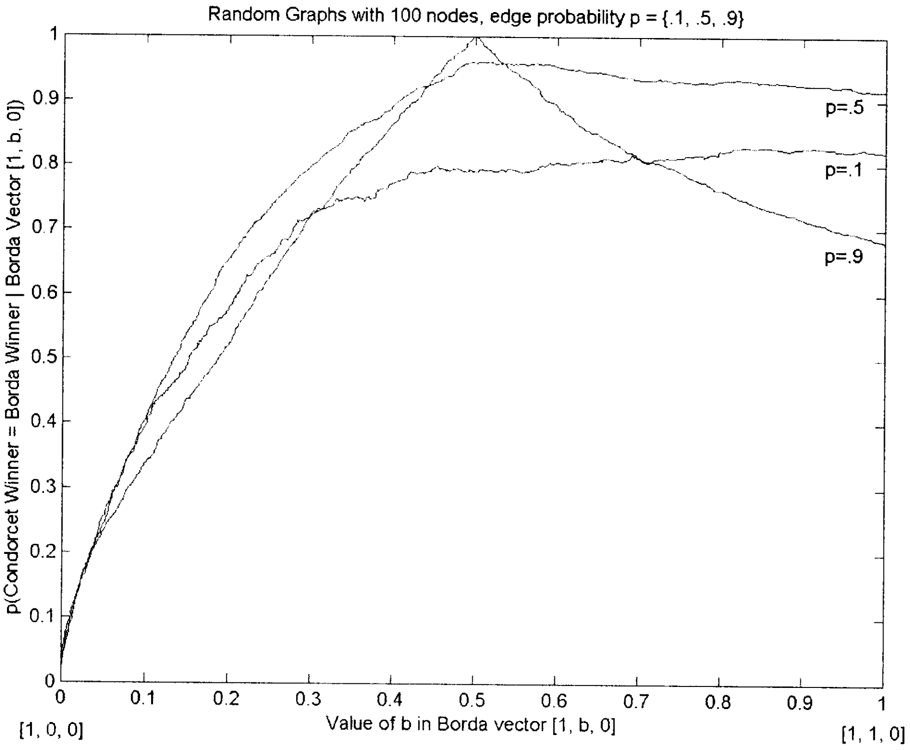

Figure 3-1: b-Borda agreement for 100 node dense random graphs

proven to be a perfect predictor of the Condorcet winner for diameter 2 graphs. Accordingly. we find that the performance of this predictor improves as n and p increase, both of which cause the expected diameter to go towards 2.

There is, however, an upper limit on the predictive power of this Borda-based method. For a graph of $n = 100, p = .5$, for example, it is not possible to achieve an accuracy higher than approximately 95%. To predict the Condorcet winner more accurately, one must consider not only the winner of the Borda tally, but rather the top k candidates, as reported by the Borda tally. Fortunately, we find that when the Borda tally does not agree with the Condorcet, it does not tend to miss too wildly. Figures 3-2 through 3-4 show the placement of the Condorcet winner in the ordering provided by the $B = \{1, .5, 0, 0, \ldots\}$ Borda tally, for a variety of n and p. We note, first of all, that the performance is very good, requiring only a handful of candidates to be considered to achieve a very high probability of success. Secondly, we note that the performance improves, once again, as n and p increase. This is a critical feature, since one must perform pairwise Condorcet tallies between the selected candidates to determine which is the actual Condorcet winner. These are not computationally cheap, so we wish to perform as few of these as possible. However, since the performance actually improves as $n \rightarrow \inf$, we can achieve an any degree of asymptotic precision by considering only a constant number of candidates. Thus, the order of growth of this is that of just a single pair-wise Condorcet tally, which, depending on the particulars of the graph structure, is often the same $O(n^2)$ as is achieved by our Borda tally.

These results, taken on the whole, tell us that it is possible to predict the Condorcet winner with arbitrary precision without resorting to a full Condorcet tally, for certain families of graphs.

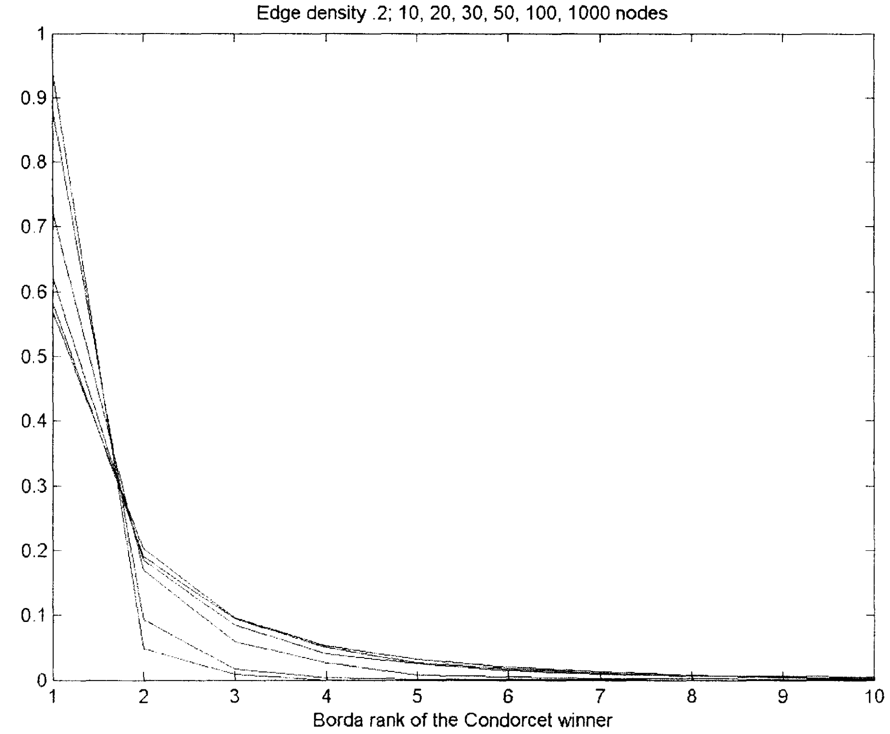

Figure 3-2: Borda Rank of the Condorcet Winner: $p = .2$

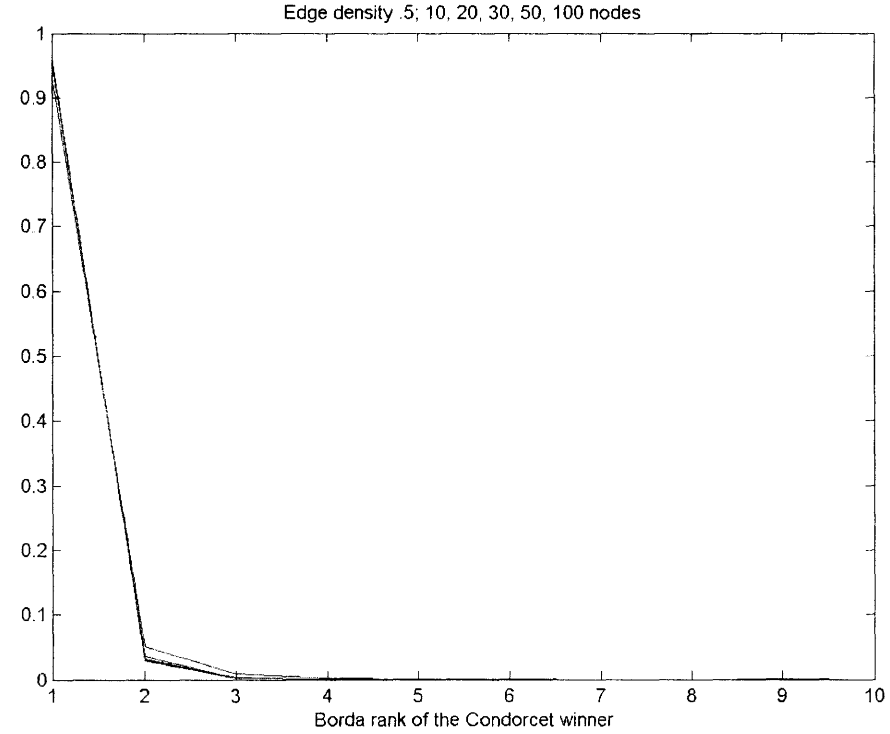

Figure 3-3: Borda Rank of the Condorcet Winner: $p = .5$

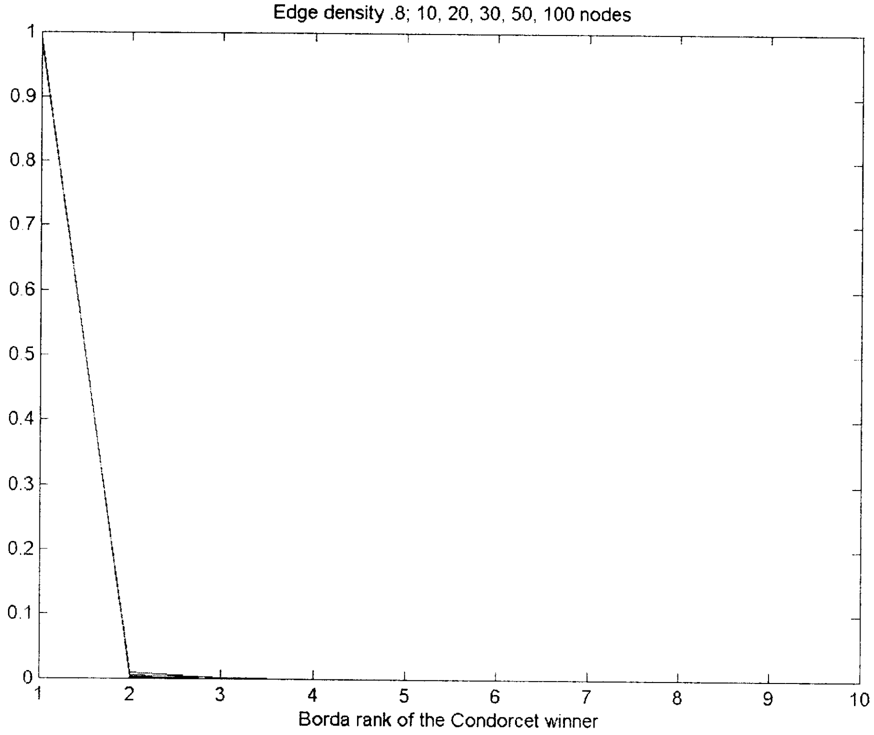

Figure 3-4: Borda Rank of the Condorcet Winner: $p = .8$

Chapter 4

Theoretical Impossibility

Unfortunately, these results are not applicable to graphs in general. In fact, there are graphs for which it is provably impossible to predict the outcome of a Condorcet tally using Borda-based methods.

4.1 Impossibility of Predicting Overall Order

First, we will perform some manipulations which will make further analysis easier. If we want to be able to predict the outcome of a Condorcet tally using one or more modified Borda tallies, we must be able to predict the outcome of a Condorcet pairwise comparison between any two candidates. Thus, we can restrict our attention to predicting the outcome of a pairwise comparison between two arbitrary nodes A and B. We note that, using the Condorcet tally, when calculating how some node C casts its points in this pairwise comparison, we only need to know the distances between A and C, and between B and C, and do not require any other knowledge of the graph structure. Thus, for a graph of diameter d, C can fall in one of d^2 distinct categories, corresponding to the possible sets of distances from the candidate nodes. We can represent these categories as a $d \times d$ matrix D, where all entries with a column are equidistant from node A, and all entries in a row are equidistant from B. We note that $D_{0,0}$ is necessarily empty if $A \neq B$, and that there is exactly one node in the leftmost column and uppermost row, these being nodes B and A, respectively.

Furthermore, if nodes A and B are some distance x from each other, then $D_{x,0} = B$ and $D_{0,x} = A$.

Within this matrix D, the upper triangular portion is closer to A than B, and the lower triangular portion is closer to B than A. Thus, we can compute the winner of this pairwise comparison by taking the sign of the sum of the piecewise product of D with the matrix

$$\begin{pmatrix} 0 & 1 & 1 & \dots \\ -1 & 0 & 1 & \dots \\ -1 & -1 & 0 & \dots \\ \vdots & \vdots & \vdots & \end{pmatrix}$$

For a 3×3 matrix corresponding to a diameter 2 graph, this computation is equivalent to

$$D_{0,1} + D_{0,2} + D_{1,2} - D_{1,0} - D_{2,0} - D_{2,1} = A + D_{1,2} - B - D_{2,1}$$

When considering the Borda tally, we can use the same matrix D. All nodes which reside at $D_{x,y}$ are distance x from A, and distance y from B, thus giving wB_x points to A and wB_y to B. Because we are only trying to compute whether $A \succ B$ or $A \prec B$ in the Borda outcome, we do not need to know the total number of points assigned to A and B, but rather simply which received more. Thus, it suffices to say that a node of weight w which resides at $D_{x,y}$ gives $wB_x - wB_y$ more points to A than to B. Thus, we can compute the winner similarly to above, except using the matrix

$$\begin{pmatrix} 0 & B_0 - B_1 & B_0 - B_2 & \dots \\ B_1 - B_0 & 0 & B_1 - B_2 & \dots \\ B_2 - B_0 & B_2 - B_1 & 0 & \dots \\ \vdots & \vdots & \vdots & \end{pmatrix}$$

4.1.1 Diameter 2 Graphs

This construction reveals a few interesting results. First, for a graph with diameter $d > 1$ (i.e. anything other than a fully-connected graph), there is no single Borda vector which yields a Borda tally that can uniquely determine the outcome of any given pair-wise Condorcet tally. Thus, we seek a more complex sufficiency condition: two or more Borda vectors whose associated Borda tallies, if in agreement, also match the outcome of the Condorcet tally.

We find that, using Borda vectors $\{110\}$ and $\{210\}$, we can do just that. That is, in a graph of diameter 2, for any pairwise comparison for which Borda tallies using the vectors $\{110\}$ and $\{210\}$ agree, the result of the Condorcet pairwise tally will be the same. To prove this, consider the two cases where A and B are distance 1 and 2 from each other separately. When they are distance 1, D takes the form

$$\begin{pmatrix} 0 & A & 0 \\ B & 0 & x \\ 0 & y & 0 \end{pmatrix}$$

so using a borda vector of $\{110\}$ produces a computation of A+x-B-y, which matches exactly the computation produced by the Condorcet tally. Similarly, when they are distance 2, D takes the form

$$\begin{pmatrix} 0 & 0 & A \\ 0 & 0 & x \\ B & y & 0 \end{pmatrix}$$

so the same outcome is reached using a Borda vector of $\{210\}$. Thus, since A and B must be either of distance 1 or distance 2, if the outcomes of using these two Borda vectors agree, then the Condorcet outcome will necessarily agree as well.

4.1.2 Diameter 3+ Graphs

This result, however, does not scale well to larger diameter graphs. In fact, we can prove that there is no set of Borda vectors whose agreement is sufficient to guarantee agreement with the Condorcet tally, if the graph's diameter is at least 3. To prove this, we consider the result of applying vector addition and scalar multiplication to Borda vectors. For some pairwise comparison, suppose we have two Borda vectors B and B' which are in agreement. This means that the $\sum D \otimes B$ and $\sum D \otimes B'$

have the same sign. It follows from the piecewise nature of the multiplication that $\sum D \otimes (B + B') = \sum D \otimes B + D \otimes B' = \sum D \otimes B + \Sigma D \otimes B'$. Thus, the Borda vector which results from the addition of two agreeing Borda vectors will also agree with those vectors. Similarly, for a Borda vector B and a positive scalar k, $\sum D \otimes kB = k \sum D \otimes B$, so kB will agree with B in all cases, if k is positive.

This means that any positive linear combination of agreeing Borda vectors will also agree with those vectors. We use this as follows: for a graph of diameter d, take the set of Borda vectors $\{100\ldots\}, \{110\ldots\}, \{111\ldots\}, \ldots$. Clearly, any valid Borda vector can be constructed via a positive linear combination of members of this set. Thus, if all members of this set agree on the result of some pairwise comparison, it necessarily follows that all possible Borda vectors would agree as well. If it were the case that these Borda vectors agreed on a result which did not agree with the Condorcet result for some pairwise comparison in some graph, then that would prove that Borda tallies are incapable of predicting that outcome.

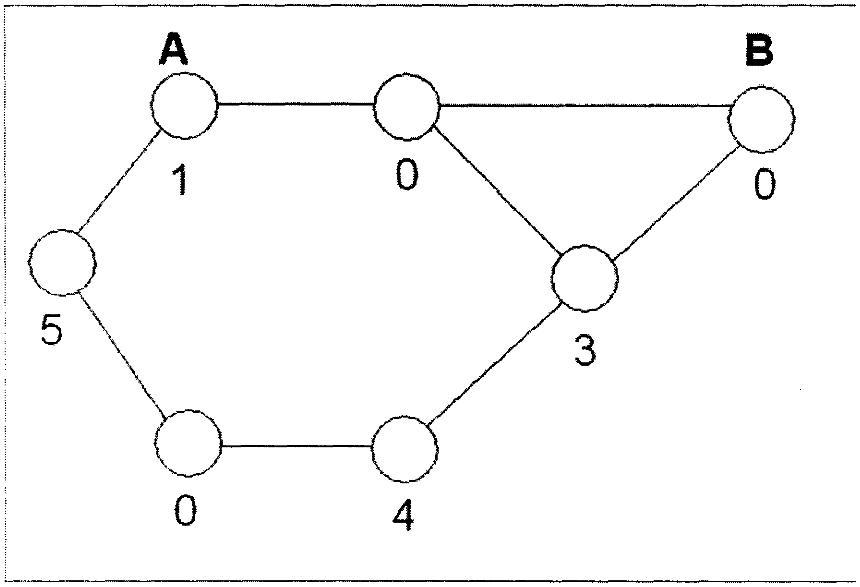

Figure 4-1: Graph for which Borda tallies cannot predict Condorcet

Consider the graph shown in Figure 4-1. Using Borda vectors of $\{1000\}$, $\{1100\}$, $\{1110\}$, $A \succ B$. However, using a Condorcet tally, $B \succ A$. Thus, it is impossible for any set of Borda vectors to properly predict the Condorcet outcome of the pairwise comparison between A and B. The Condorcet tally produces a transitive outcome for this graph, and it is trivially possible to embed this graph in a larger diameter graph with the same set of properties. Thus, there is a class of graphs for which the goal of using a modified Borda tally to predict Condorcet in those cases in which Condorcet is transitive is infeasible, for graphs of diameter 3 or greater.

4.2 Impossibility of Predicting Winners

An astute reader might note, at this point, that the reversal present in the graph shown in Figure 4-1 occurs in the middle of the Borda and Condorcet orders; neither the A nor B in question is the Condorcet winner of that graph. Especially given the focus of earlier sections on agreement of the winner, it would be reasonable to then ask whether a similar impossibility exists for the winner - that is, whether it might be possible for Borda-based methods to predict the Condorcet winner, if not the entire order. Unfortunately, this is also impossible. We show this again by constructing a graph for which there is a pair of nodes A and B such that B is the Condorcet winner (and thus, $B \succ A$ in the Condorcet tally, as above), yet $A \succ B$ in any Borda tally. This time, the graph is neither simple nor planar, so rather than just presenting it, we describe its construction. Consider a graph G exhibiting the reversal demonstrated in section 4.1.2, for example that shown in Figure 4-1. We divide this graph into subgraphs: node A, node B, and $G' = G - A - B$. Now, we make k copies of G': $G'_1 \ldots G'_k$, where $k \gg n$. These are connected internally in a manner identically to G', and every node in G'_i corresponding to a node that was connected to A or B in G is similarly connected to A or B. This graph, in essence, is the result of replicating by k times all nodes other than A and B, then connecting them back in a manner consistent with each node's respective position in G.

Now, we consider the Condorcet winner of this graph. The winner will clearly be either A or B, since every member of G'_i is at least as close to A and B as it is to any member of $G'_{j \neq i}$, and thus cannot prefer any node in a foreign G' to either A or B. Since $B \succ A$ in the original graph G, and the shortest paths to A and B

from within any G' were not affected by the additions of the other G's, it follows that $B \succ A$ in the Condorcet order for this new graph, as well. Thus, B is the Condorcet winner for the new graph. By a similar argument, we had found that A accrued more points than B in any Borda tally in the original graph. Since we are, once again, not modifying the shortest path from any member of G' to A or B, there is no way for the outcome of any Borda tally on this new graph to be any different, so it follows that $A \succ B$ in any Borda tally. Thus, we have constructed a graph for which no Borda tally can predict the Condorcet winner, by an argument paralleling that from section 4.1.2.

Chapter 5

Application to Small-World Graphs

In this section, we present some results relating to another class of interesting graphs: small-world graphs[9]. These graphs are of interest primarily because of their practical relevance, since they are often used as representations of "complex systems." The general notion of a small-world graph is that it represents the "small world" phenomenon popularized by a game surrounding actor Kevin Bacon. One can select nearly anyone associated with the film industry, and by following links between actors formed by mutual appearance in a film, trace a path to Kevin Bacon is usually no more than 6 steps. However, the graph of links among actors is not particularly dense, so it might seem surprising that its diameter is apparently so small. This is the small world property, which can be formalized as follows.

We define two metrics for a graph, the Characteristic Path Length (CPL) and the Clustering Coefficient (CC). The CPL of a graph is simply the mean of the lengths of the shortest paths between all pairs of nodes. Or, in mathematical notation,

$$CPL(G) = \frac{\sum_{i \in G, j \in G, i \neq j} d_{i,j}}{|G|(|G|-1)}$$

The CC of a graph is, directly, the fraction of mutual neighbors of any given node that are connected to each other, and is often taken as a representation of how tightly clustered a graph is. Mathematically, we define

$$CC(G) = \frac{\sum_{i\in G} \frac{\sum_{j\in G,k\in G} E_{i,j}E_{i,k}E_{j,k}}{\sum_{j\in G,k\in G} E_{i,j}E_{i,k}}}{|G|}$$

where $E_{i,j}$ is defined as 1 if i and j are connected in G, and 0 otherwise.

Small-world graphs, then, are graphs which have a low CPL but a high CC, where those "low" and "high" are with respect to standards which vary based on the domain. One method for generating small-world graphs is known as "rewiring." This method involves starting with a regular graph which we call a "k-braid," (where k is the vertex degree of the regular graph) which consists of nodes indexed such that each is connected to all nodes whose index is within $\frac{k}{2}$ of its own, and is not connected to any other nodes. From this braid, each edge has one of its vertices reassigned to a random node in the graph with probability p - this is the "rewiring." For $p = 0$, this simply yields a braid. For $p = 1$, this results in a traditional random graph with edge density $\frac{k}{n}$. As p is increased between 0 and 1, the CPL drops much more quickly than the CC, yielding a range of values for p which result in small-world graphs.

This domain is interesting in and of itself. Generating such a graph, we can discuss the probability that it manifests a unique Condorcet winner, as a probability taken over the domain of all possible weight assignments to its nodes. The graph structure specifies the connections between the nodes, but the values of the weights can be taken as a point in a n-dimensional hypercube. Clearly, there are weight assignments which will always produce a Condorcet winner, such as assigning a 1 to a single node, and 0 to all other nodes. Thus, we know that $p(CondorcetWinner) > 0$. However, in most circumstances, there exist assignments which cause the Condorcet tally to produce invalid, non-transitive orderings. So, for any given graph, we can calculate the probability of there being a unique Condorcet winner, which will result in a number in the range of $(0, 1]$. This probability is incredibly difficult to calculate directly for large graphs, so we use Monte Carlo methods to estimate it. Figure 5-1 shows the distribution of this probability for small-world graphs formed by taking a 10-braid of size 1000 and rewiring it with probability $p = .005$.

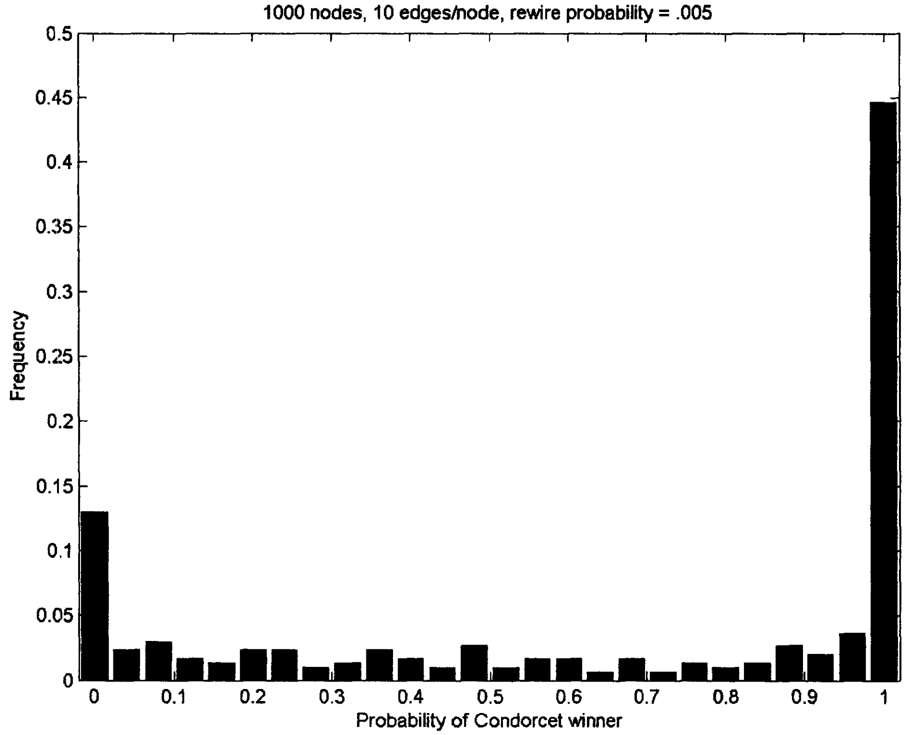

Figure 5-1: Distribution of $p(CondorcetWinner)$ for small-world graphs

As can be seen from this graph, there is a very strong tendency for these graphs to have a very high ($> 95\%$) probability of yielding a unique winner, doing so approximately 45% of the time, but that they also have a tendency towards having a very low ($< 5\%$) probability of yielding a winner, doing that approximately 15% of the time. This type of strongly bi-modal distribution suggests that, at least with respect to the Condorcet properties of a graph, there are two distinct sub-types of small-world graphs. Condorcet probability characteristics can be used to describe the "stability" of a graph, in terms of its reaction to small perturbations, so, given that these graphs are often seen in practice as descriptions of complex systems, such a result is potentially very interesting.

5.1 Random Regular Graphs

As a related experiment, we investigate a slightly modified method for constructing a small-world graph. Since we rewire randomly, we change the degrees of the vertices each time we rewire, so, although we started with a regular graph, the resulting graph is not regular. Via a novel modification to the rewiring process, however, we can maintain regularity throughout the process. This results in random regular graphs, wherein each node is locally identical, being connected to the same number of neighbors. These graphs are potentially more interesting for their lack of natural centers (a la Kevin Bacon), which tend to have higher than average degrees. We can create these graphs through a process we call rewire-repair. The process of rewiring can be seen as selecting an edge which connects the **location** and the **donor**. The location's degree remains unchanged, but the donor loses an edge. We then connect this edge to the **acceptor**, whose degree then increases by 1. We then fix this via repair. To repair, we select the **acceptor-donor**, which is yet another node which is connected to the **acceptor**. We sever the connection between the **acceptor-donor** and the **acceptor**, returning the **acceptor**'s degree to its former value. We then connect the **acceptor-donor** to the **donor**, increasing the latter's degree so that it, too, regains its former value. Thus, all nodes end up with the same degree after the rewire-repair step as they had before, so it does not change the regularity of the graph. We find that performing rewire-repair with probability p has a very similar effect on the graph's CPL and CC as performing simple rewire, which leads us to the conclusion that we can thus create random regular small-world graphs. Figure 5-2 shows this, plotting CPL and CC as a function of p for both standard rewire and rewire-repair.

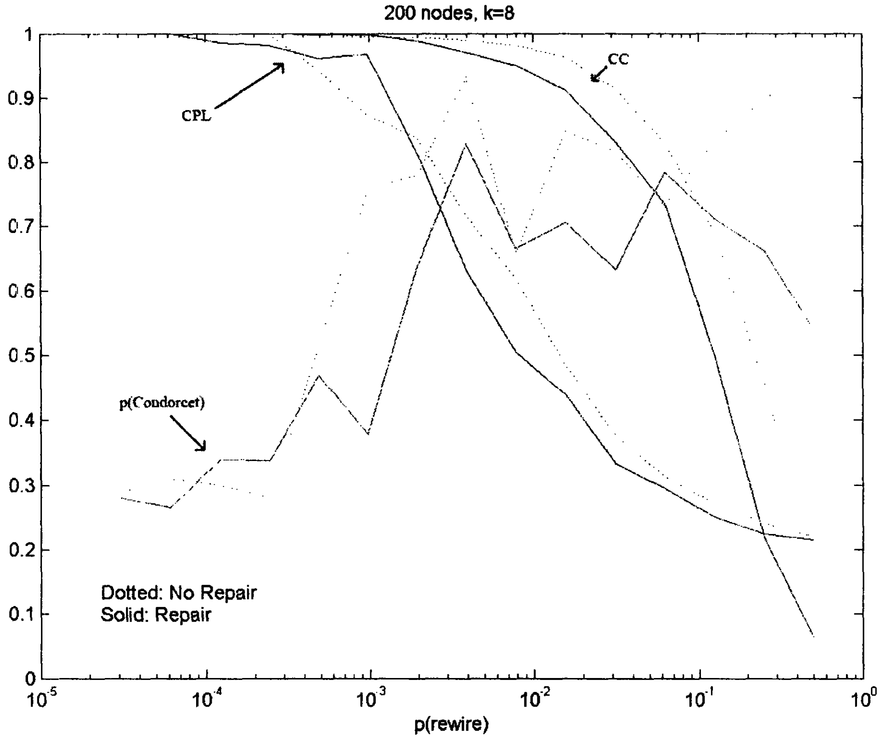

Figure 5-2: CPL, CC, and $p(CondorcetWinner)$ as a function of p, for standard rewire and rewire-repair

Figure 5-2 shows something else which is perhaps more interesting, however. As $p \rightarrow 1$, $p(CondorcetWinner) \rightarrow 1$ for standard rewire. This fits well with previous results, in which random graphs were shown to have very high probability of yielding a unique Condorcet winner[7]. On the other hand, rewire-repair deviates significantly at that point, approaching a value of $p(CondorcetWinner) \approx \frac{1}{2}$ as $p \rightarrow 1$. Investigating this more closely, we can plot the distribution of $p(CondorcetWinner)$ at $p = 1$. The result, shown in Figure 5-3, is surprising: each instance has a $p(CondorcetWinner) \approx$

$.5 \pm .1$.

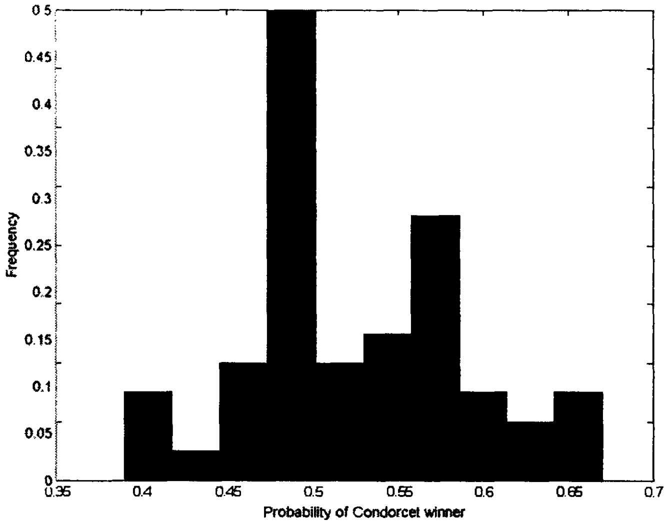

Figure 5-3: Distribution of $p(CondorcetWinner)$ for random regular graphs, $k = 10$, $n = 1000$

This is interesting for a number of reasons. Firstly, it is simply abnormal for graphs to have a distribution of $p(CondorcetWinner)$ clustered so strongly around a value other than 0 or 1. Secondly, this is a relatively natural class of graphs, which is potentially relevant to some computer networking problems in which computers are connected to a random set of peers (distributed file sharing, etc.). This Condorcet characteristic is indicative of incredible instability, as the information-theoretic properties of the graph depend heavily on the set of values assigned to the nodes. The ramifications of this observation, however, are beyond the scope of this thesis, and more work is needed to fully explore them.

Chapter 6

Conclusion

In this thesis, we presented a number of information-theoretic results relating to information aggregation in domains with shared knowledge structures which can be represented in the form of graphs. We presented numerical results showing that there exist very good predictors between relevant aggregation methods for certain classes of graphs, but also that such predictors are provably guaranteed to make errors on some classes of graphs. Thus, we demonstrated that the goal of a Borda-based method which satisfies the Condorcet criterion whenever possible is, in fact, unachievable. Finally, we presented some related work demonstrating some potentially very practically relevant properties of graphs which are often used in practice to represent complex systems. This last result, in particular, opens more doors than it closes, and it is our hope that continued research in this domain will prove fruitful to the end of more fully characterizing the properties of these families of graphs.

Bibliography

[1] Kenneth Arrow. *Social Choice and Individual Values.* Yale University Press, New Haven, CT, 2 edition, 1963.

[2] Jean-Charles de Borda. *Memoire sur les elections au Scrutin.* Histoire de l'Academie Royal des Sciences, 1786.

[3] Marquis de Condorcet. *Essai sur l'application de l'analyse a la probabilite des decisions rendues a la probabilitie des voix.* De l'imprinerie royale, Paris, 1785.

[4] Diana Richards. Coordination and shared mental models. *American Journal of Political Science*, 45:259–276, 2001.

[5] Whitman Richards. Anigrafs: Experiments in collective consciousness. http://people.csail.mit.edu/whit/contents.html.

[6] Whitman Richards, H. Sebastian Seung, and Galen Pickard. Neural voting machines. Publication Forthcoming, May 2006.

[7] Whitman Richards, Nicholas Wormald, and Galen Pickard. Asymptotes for condorcet winners. Publication Forthcoming.

[8] Daniel Saari. *Geometry of Voting.* Springer-Verlag, New York, 1995.

[9] Duncan Watts and Steven Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:493–527, 1998.

[10] Peyton Young. Optimal voting rules. *Journal of Economic Perspectives*, 9:51–64, 1995.