

# RESEARCH STATEMENT

STEPHEN J. YOUNG

When I first heard about Georgia Tech’s program in Algorithms, Combinatorics and Optimization (ACO), I knew I had found a program that fit my interests ideally. Although I considered myself primarily a mathematician, my interests ranged over a variety of topics that are not necessarily found in every mathematics department, such as the foundations of computer science, operations research, algorithm design, etc. The ACO program seemed tailor-made for my interests, focusing on the interface and interplay between mathematics, computer science, and operations research. As a result of being in the ACO program I have had the opportunity to work with a variety of people in a variety of fields at Georgia Tech. The interdisciplinary nature of the ACO program is reflected in the nature of my current research projects. Over the next few years I see this multidisciplinary focus continuing, as I see myself continuing my current research in mathematical modeling, combinatorial structures, especially partially ordered sets, and complex networks.

## 1. MATHEMATICAL MODELING

Throughout my undergraduate career the application of mathematics, especially discrete mathematics, to the solving of real world problems fascinated me. In every year of my four year career at Rose-Hulman, I competed in the Mathematical Contest in Modeling. Although the teams I was on never produced one of the few winning papers, we performed well each year, receiving three meritorious ratings and one honorable mention. Even though mathematical modeling is not my primary focus, my interest in these interdisciplinary and collaborative problems continues. For instance, I have recently begun a project with Prof. Joel Sokol of the School of Industrial and Systems Engineering at Georgia Tech, focusing on the manner in which a professional baseball team should choose whom to draft. Because of the extensive nature of the professional baseball draft (30 teams and approximately 50 rounds) teams need to scout roughly 1800 – 2000 players. It is obvious that no one scout, or small collection of scouts, can observe all 2000 players. Thus, the team is left with the task of collating the scouting reports of approximately 20 scouts, each of whom has varying scouting abilities and has seen a relatively small fraction of the players. Furthermore, there are a limited number of players who have been scouted by multiple scouts. We are currently experimenting with several methods of aggregation to find a total ordering of the scouted players that is mostly “right.” The goal of our modeling is not to replace the “discussions” the team will have regarding the draft order, but to narrow down the discussions to the most contentious areas.

## 2. COMBINATORIAL STRUCTURES

*Linear Discrepancy* The linear discrepancy of a partially ordered set (poset) is one way of measuring the “distance” a poset is from a linear order. Given a poset  $P$  and linear extension of the poset  $L$ , the uncertainty of the linear extension is defined as  $\max_{x||y \in P} |h_L(y) - h_L(x)|$  where  $h_L(x)$  is the height of  $x$  in the linear extension  $L$ . Then the linear discrepancy of a poset,  $ld(P)$ , is defined as the minimum of the uncertainty over all linear extensions of the poset.

The first natural question regarding linear discrepancy is whether there is some characterization of posets with a given linear discrepancy. It is clear that if the linear discrepancy is 0 then the poset is a chain. In [13], Tanenbaum, Trenk and Fishburn show that the posets with linear discrepancy 1 are precisely the semi-orders of width 2. Chae, Cheong, and Kim reformulated the characterization of posets with linear discrepancy 1 in terms of minimal posets with linear discrepancy 2. They refer

to a minimal poset with linear discrepancy  $k$  as a  $k$ -discrepancy-irreducible poset [1]. Rautenbach then conjectured that there is a finite list of 3-discrepancy-irreducible posets [11]. This was proven to be false by Howard in [5], where he shows that there is an infinite family of 3-discrepancy-irreducible posets of width 2. Together with Howard and Keller, I completed the characterization of posets of linear discrepancy 2 by finding an infinite family of 3-discrepancy-irreducible posets of width 3 and proving that, together with the list of Howard, all 3-discrepancy irreducible posets had been described [6]. Since then, we have computationally explored the posets of linear discrepancy 4 and we believe that, as in the case of dimension, the question of irreducibility becomes significantly more complex for values greater than 3. Since then, I have shown that within any poset of linear discrepancy at least  $k$ , there is a  $k$ -discrepancy-irreducible poset, and thus a characterization of  $k$ -discrepancy-irreducible posets provides a forbidden subposet characterization of linear discrepancy at most  $k - 1$ . The key observation in the most recent of these results is that the uncertainty of a linear extension is determined entirely by the behavior of the critical pairs rather than all incomparable pairs. This observation allowed us to relate the 3-discrepancy-irreducible posets of width 3 to the previously discovered 3-discrepancy-irreducible posets of width 2 via a “filling in” of select critical pairs. I believe that this operation generalizes to higher linear discrepancy. That is, I believe that given a  $k$ -discrepancy-irreducible poset containing a critical pair of a certain unknown form, the critical pair can be “filled in” and the resulting poset will still be  $k$ -discrepancy-irreducible. My first step in attempting to determine the form of these critical pairs will be to examine which critical pairs have this property for 4-discrepancy-irreducible posets. My long term goal is to use the insight gained from studying the relationship of critical pairs and linear discrepancy to resolve the conjecture of Tanenbaum, Trenk, and Fishburn that, if  $\text{ld}(P) = \dim(P) \geq 5$ , then  $\mathbf{S}_{\text{ld}(P)} \subseteq P$  [13].

In [2], Tanenbaum, Trenk and Fishburn show that the linear discrepancy of a poset is equal to the bandwidth of the co-comparability graph. Letting  $\Delta_P$  be the maximum degree in the co-comparability graph of  $P$ , Trenk, Tanenbaum and Fishburn conjecture that  $\text{ld}(P) \leq \lfloor \frac{3\Delta_P - 1}{2} \rfloor$ . This conjecture is motivated by the observation that the linear discrepancy of the disjoint sum of two chains of length  $\Delta$  is  $\lfloor \frac{3\Delta - 1}{2} \rfloor$ , and so if it holds it is tight for all  $\Delta$ . Rautenbach shows a partial result in this direction by proving that  $\text{ld}(P) \leq 2\Delta_P - 2$ . [11] Together with Keller, I have shown that for interval orders,  $\text{ld}(P) \leq \Delta_P$ , and consequently the bandwidth of an interval graph is at most its maximum degree [7]. Trotter, [15], has exhibited a family of width 3 interval orders such that this bound is tight in the limit. Independently, I have shown that if there exists a counterexample to the conjecture of Tanenbaum, Trenk and Fishburn, then the minimal such counterexample is a connected poset. The natural further direction is resolve whether the conjectured upper bound holds and whether it is tight for connected discrepancy-irreducible posets. Based on empirical observations, I believe that for discrepancy-irreducible posets, a tighter bound may hold for connected posets which would in turn imply that Tanenbaum, Trenk, Fishburn conjecture is true and is tight.

### 3. MODELS FOR COMPLEX NETWORKS

I was first introduced to the random dot product graph model by E.R. Scheinerman through his work with Kraetzl and Nickel. The model, as they had formulated, was a three parameter model,  $(\alpha, d, n)$ . Each of  $n$  vertices was randomly assigned a  $d$ -dimensional vector, with each component distributed independently as  $\frac{1}{\sqrt{d}}\mathcal{U}^\alpha[0, 1]$ . Then each edge was present, independently, with probability equal to the the dot product of the vectors assigned to its endpoints. They show in [10], that with  $d = 1$ , their model for random dot product graphs has the following properties:

$$(1) \left( \frac{\alpha+1}{2\alpha+1} \right)^2 = \mathbb{P}(u \sim w \mid u \sim v \sim w) > \mathbb{P}(u \sim w) = (\alpha + 1)^{-2}.$$

- (2) If  $\lambda(k)$  is the random variable indicating the number of vertices of degree  $k$ , then  $\mathbb{E}[\lambda(k)] \sim \frac{1}{k!^\alpha} (1 + \alpha)^{\frac{1}{\alpha}} \Gamma\left(\frac{1}{\alpha} + k\right) n^{\frac{\alpha-1}{\alpha}}$  for  $k \in \mathbb{Z}^+$  as  $n \rightarrow \infty$ .
- (3) The giant component has diameter at most 6 as  $n \rightarrow \infty$ .

Working with Scheinerman, I was able to generalize their model to allow the vectors to distributed according to any distribution that satisfies the condition that the dot product of two samples is almost surely in the interval  $(0, 1)$ . In this more general model, we were able to show that  $\mathbb{P}(u \sim w \mid u \sim v \sim w) \geq \mathbb{P}(u \sim w)$  and that an arbitrarily large fraction of the graph is connected with diameter at most 5 as  $n \rightarrow \infty$ . Further, we were able to show that there is an integral formula for  $\mathbb{E}[\lambda(k)]$  that is the expectation of a continuous function of the random variable depending on  $n$ ,  $k$ , and the expectation of distribution. With further work, Scheinerman and I were able to generalize this model to a directed model, with similar results. Specifically, an arbitrarily large fraction of the graph is strongly connected with diameter 5, the graph exhibits a directed version of clustering, and the expected in-degree and out-degree can be calculated as the expectation of a function of the random vectors [12].

Given a cut,  $(S, \bar{S})$ , in a graph, define  $C(S, \bar{S})$  to be number of edges crossing the cut and  $\text{Vol}(S)$  be the number of edges in the graph induced by  $S$ . Then the conductance of a graph is defined as the minimum of  $\frac{C(S, \bar{S})}{\min\{\text{Vol}(S), \text{Vol}(\bar{S})\}}$  over all cuts in the graph. The conductance of a graph has important algorithmic implications. In particular, conductance that is constant in the number of vertices has been shown to imply “good” behavior of search algorithms, network congestion, etc. [3, 4]. Although I have not yet been able to show that the random dot product graph has constant conductance in general, I have been able to show that by restricting to “semantic/geometric” cuts, the model has constant conductance. That is, if the only cuts under consideration are those cuts generated by a partition of the underlying space, then the conductance is constant as a function of  $n$ .

One obstacle to the practical application of the random dot product graph model is that the average degree of a vertex is on the order of the number of vertices, as opposed to the sublinear average degrees in observed in many networks of interest. However, recently I have been able to modify the model to allow logarithmic average degree with only minor adjustments to the known results. In particular, all the known results hold except that an arbitrarily large fraction of the graph is now connected with diameter at most  $2 \log(n) + 1$ .

This leaves open several natural questions which I will continue to work on in the future. First, from a modeling point of view, it would be desirable to have a means of generating a specific degree sequence. Specifically, given a degree distribution (directed or undirected) is there a way to construct a distribution so that the expected degree sequence approximately matches the given degree sequence? More generally, given the flexibility of the model, what parameters of the graph can be reverse engineered? For instance, would it be possible to specify the density of certain subgraphs?

I am sure that the real potential of the random dot product graph model lies, however, in its broad range of potential applications. The key to these applications is that we can view the vector associated with a vertex as “encoding” semantic content about the vertex. This semantic interpretation leads me to believe that, similarly to Kleinberg’s work [9], there should be a way to leverage the semantics to show that there is efficient navigation within the network under some loose conditions. Perhaps more practically interesting, I would like to study the nature of virus propagation in the random dot product graph model. A result on virus propagation may give insight into new inoculation schemes exploiting semantic information, which could limit the spread of viruses in certain situations, whether they be computer viruses over the Internet or biological viruses.

#### 4. FUTURE DIRECTIONS

Over my career I envision myself maintaining my multidisciplinary focus within the broad field of discrete mathematics. For example, some of the questions I intend on working in the long term:

- ◆ *Dimension of the Poset Product* Currently not much is known about the behavior of dimension under the operation of poset product. It is obvious that  $\max\{\dim(P), \dim(Q)\} \leq \dim(P \times Q)$  and with a little work it can be shown that  $\dim(P \times Q) \leq \dim(P) + \dim(Q)$ . Trotter showed that  $\dim(S_n \times S_m) = n + m - 2$  [14], but this leaves open many questions. Can the lower or upper bound be improved? Is there a poset  $P$  so that  $\dim(P \times P) = \dim(P)$ ? Are there structural characteristics of  $P$  and  $Q$  that determine the relationship between  $\dim(P)$ ,  $\dim(Q)$  and  $\dim(P \times Q)$ .
- ◆ *Performance of First Fit Graph Coloring* Consider the problem of coloring an interval graph with clique number  $k$ . One natural question is to determine the worst case behavior of the greedy algorithm on this problem. The first upper bound on the worst case performance was provided by Woodall in 1976, followed twelve years later by the first linear bound of  $40k$  by Kierstead [8]. Currently the best known results are an upper bound of  $8k - 3$  and a lower bound of  $(5 - \epsilon)k$  for any  $\epsilon > 0$  and sufficiently larger  $k$ . Although there is not much room for improvement here, based on the methods of proof used so far, I believe there is an opportunity for a clever idea to resolve this issue in its entirety.
- ◆ *Biological Combinatorics* This last area is more of a general field. With the increasing amount of biological data available, from sources like the Human Genome project, it is apparent that mathematics is going to play a key role in efficiently analyzing and interpreting this data. In addition, mathematics is seemingly playing a greater role in biological research, especially in the field of protein folding. Although I do not have much knowledge of these fields at this moment, I am looking forward to the opportunity to learn how discrete mathematics can answer some of the vital questions in biology.

#### REFERENCES

- [1] G.-B. CHAE, M. CHEONG, AND S.-M. KIM, *Irreducible posets of linear discrepancy 1 and 2*, Far East J. Math. Sci. (FJMS), 22 (2006), pp. 217–226.
- [2] P. C. FISHBURN, P. J. TANENBAUM, AND A. N. TRENK, *Linear discrepancy and bandwidth*, Order, 18 (2001), pp. 237–245.
- [3] C. GKANTSIDIS, M. MIHAIL, AND A. SABERI, *Conductance and congestion in power law graphs*, SIGMETRICS Perform. Eval. Rev., 31 (2003), pp. 148–159.
- [4] ———, *Hybrid search schemes for unstructured peer-to-peer networks*, INFOCOM, (2005).
- [5] D. M. HOWARD, *3-discrepancy-irreducible posets of width 2*. Preliminary Manuscript, 2007.
- [6] D. M. HOWARD, M. T. KELLER, AND S. J. YOUNG, *A characterization of partially ordered sets with linear discrepancy equal to 2*, Order, 24 (2007), pp. 139 – 153.
- [7] M. T. KELLER AND S. J. YOUNG, *The linear discrepancy of interval orders*. In Preparation, 2007.
- [8] H. A. KIERSTEAD, *The linearity of first-fit coloring of interval graphs*, SIAM J. Discrete Math., 1 (1988), pp. 526–530.
- [9] J. M. KLEINBERG, *The small world phenomenon: an algorithmic perspective*, in STOC '99: Proceedings of the thirty-second ACM Symposium on the Theory of Computer Science, 1999.
- [10] M. KRAETZL, C. NICKEL, AND E. R. SCHEINERMAN, *Random dot product graphs: A model for social networks*. Preliminary Manuscript, 2005.
- [11] D. RAUTENBACH, *A note on linear discrepancy and bandwidth*, J. Combin. Math. Combin. Comput., 55 (2005), pp. 199–208.
- [12] E. R. SCHEINERMAN AND S. J. YOUNG, *Directed random dot product graphs*. preliminary manuscript, 2005.
- [13] P. J. TANENBAUM, A. N. TRENK, AND P. C. FISHBURN, *Linear discrepancy and weak discrepancy of partially ordered sets*, Order, 18 (2001), pp. 201–225.
- [14] W. T. TROTTER, JR., *The dimension of the Cartesian product of partial orders*, Discrete Math., 53 (1985), pp. 255–263. Special volume on ordered sets and their applications (L'Arbresle, 1982).
- [15] W. T. TROTTER, JR. personal communication, October 2007.