

Research Statement

Manan Sanghi

Department of Electrical Engineering and Computer Science
Northwestern University
Evanston, IL, USA

My research interests center design and analysis of algorithms and span multiple application areas including bioinformatics, network security, DNA word design, musical information retrieval, autonomic computing, P2P networks and e-commerce. In addition to research in algorithms for combinatorial optimization, I have a track record for collaborating with different research groups and designing solutions to computational problems they encounter.

The first and perhaps the greatest challenge in interdisciplinary research is the formulation of the correct computational problem. Often we are faced with computational problems that are NP-Hard. Therefore, under the widely believed conjecture that $P \neq NP$, no polynomial time algorithm can solve them in the worst case. Fortunately, it is usually possible to devise heuristics that give good performance in practice in spite of negative theoretical results. To me this indicates that a worst case analysis is perhaps too prohibitive. Average case analysis on the other hand could be unconvincing for it involves assuming the input to be drawn from some random distribution. *Existence of such algorithms which work well in practice but evade traditional analysis techniques is what intrigues me the most.*

I had some success in the design and analysis of one such algorithm for a network security problem recently which I describe next. I then discuss a recent work on DNA word design, an interdisciplinary research project and conclude with some fundamental research directions.

1 Automatic Signature Generation for Polymorphic Worms

Worms are self-propagating computer programs that threaten to compromise all vulnerable machines over the internet in a matter of few hours or even minutes. Due to this fast propagation, manual generation of signatures is unviable for computer worms. To further complicate the problem, worms could be polymorphic. That is they could change their byte sequence at every successful infection to evade detection using signatures.

In collaboration with the network security group at Northwestern we are developing a system named *Hamsa* for automatic signature generation for polymorphic worms [4]. Hamsa is based on the assumption that a worm must exploit one or more server specific vulnerabilities. This constrains the worm author to include some invariant tokens $\mathcal{I}$ that are crucial for exploiting the vulnerability. We formally capture this idea by means of an adversary model which allows the worm author to include any byte strings for the worm flows as long as each flow contains tokens present in the invariant set $\mathcal{I}$.

Like most practical optimization problems, the most natural computational problem for this case also turns out to be NP-Hard. However, we design an algorithm which not only performs very well in practice but also achieves provable performance guarantee by a model based analysis. Though the problem is hard in the worst case, under certain uniqueness assumption on the invariant tokens $\mathcal{I}$ we are able to guarantee signatures with bounded false positives and false negatives. From the security perspective, this is perhaps the first work with some form of provable attack resilience.

Future Directions Hamsa is currently a pure pattern based signature generator. It is widely believed that polymorphic worms of the future may be immune to such pattern-based approaches and hence the signatures should exploit some semantic information. We are currently exploring different approaches to generate a hybrid of semantic and pattern based signatures.

2 Graph Labeling and DNA Word Design

Given an input graph G , we consider the problem of labeling its vertices with equal length binary strings of minimal length such that the Hamming distance is small between words of adjacent nodes and large between words of non-adjacent nodes. The problem is motivated by growing applications in DNA computing and DNA self-assembly.

Early applications in the field have required building a set of n equal length code words such that hybridization between the words or Watson Crick complement of words is unlikely. Using hamming distance to approximate how well two words bind, such a requirement translates to designing a set of n words such that the Hamming distance between any pair in the set is large.

Basic DNA Word Design In [2], we consider a wide variety of combinatorial constraints beyond the hamming distance constraint to capture the non-interaction and stability requirements of the DNA strands as proposed by biologists. Our algorithms for the purpose employ randomization techniques and fast fourier transforms to efficiently solve the problem for a large number of constraint subsets in polynomial time while achieving a word length within a constant factor of the optimal. Our results offer a strong theoretical base for DNA word design problems, in contrast to previous approaches that are heuristic in nature.

Flexible Word Design For some recent applications, hybridization between different words in a code word set is desirable and necessary. That is, there is growing need for the efficient design of DNA codes such that the hybridization between any two words in the code is determined by a graph specifying which strands should bond and which should not. For general graphs we have designed algorithms that bound the word length with respect to either the maximum degree of any vertex or the number of edges in either the input graph or its complement in [1]. We further provide multiple types of recursive, deterministic algorithms for trees and forests, and an improvement for forests that makes use of randomization. We consider generalizations of this problem to weighted graphs and graphs with optional edges and also explore the extension from simple adjacency queries to more general distance

queries. We show how to obtain distance labelings for rings and paths by applying properties of hypercube traversal.

Future Directions Kao et al. [2] discusses nine different combinatorial constraints and proposes polynomial time algorithms achieving polynomial word length for a variety of subset of them. To be able to do so for all nine constraints simultaneously is still open.

The graph labeling problem introduced in [1] opens up avenue for many exciting research directions. For different special classes of graphs like planar graphs and chordal graphs, it may be possible to achieve a much shorter length of labeling than that for general graphs. Furthermore, there is a gap between the information theoretic lower bound on the length of labelings and the lengths achieved for the different cases considered in [1].

3 Algorithms, Biology and Music

There are few things more satisfying than seeing how ideas from one field inspire solutions in another.

Typically search for music is done using meta-information like the name of the song, the artist, the album and so on. Query by Humming systems are content based music information retrieval systems that allows the user to find a song by humming part of the tune. This can be achieved by representing songs in the database and the hummed query as strings over an alphabet of notes. Matching the query to a song then is very similar to the string alignment problem biologists face when searching for homologues of a newly discovered gene or protein sequence in an existing database.

However, there are a few differences. While in biology both the query and the target are single sequences, in the case of music the songs in the database could be polyphonic with different tracks playing in parallel which are best represented as a set of multiple sequences rather than a single sequence. The hummed query could jump from one track to another and hence need to be aligned with all the tracks simultaneously. Multiple sequence alignment is known to be a very hard problem in computational biology. However, we soon realized that the multiple sequences corresponding to a musical score are in a sense already aligned and therefore allow for a polynomial time dynamic programming solution discussed in [3].

Future Directions Interestingly, the algorithms we developed for musical information retrieval may inspire novel solutions to applications back in computational biology! Given a set of aligned sequences belonging to a particular family of DNA or protein sequences and a newly discovered sequence, it is often of interest to determine whether the new sequence belongs to that family. Typically this is done by comparing the new sequence with the consensus of the aligned sequences. However, it is conceivable that in some cases a member of the same family could ‘jump’ from one sequence to another in the family (like a hummed query jumps from one track to another) instead of matching some common consensus. Obviously there are important differences but this work inspires exploring a computational biology problem in a new light.

4 Fundamental Research Directions

Besides the goals I have for the individual projects (for a more exhaustive list of my recent research projects, please see www.cs.northwestern.edu/~msanghi/research), my vision is to develop novel techniques for analyzing algorithms in two fundamental ways.

1. Analyzing the problem complexity in terms of input distribution instead of input size alone. I was introduced to the idea of distribution sensitive analysis by my undergraduate advisor Prof. Sandeep Sen (at IIT, Delhi) [5] and it was the subject of my undergraduate research project.
2. Model-based analysis of algorithms which deals with explaining the performance of algorithms that work well in practice but are not amenable to traditional worst-case or average case analysis. I had some success in this direction for the signature generation problem for polymorphic worms discussed earlier.

In particular, I would like to employ model based analysis for different motif-finding and data mining applications. It could also prove useful for analyzing performance of recommendation systems. Besides this, I have been studying the game theoretic ideas developed by economists and would like to explore their potential applications in design of mechanisms for distributed autonomous systems and reputation systems.

References

- [1] M.-Y. KAO, M. SANGHI, AND R. SCHWELLER, *Flexible Word Design and Graph Labeling*, under submission for journal publication.
- [2] ———, *Randomized Fast Design of Short DNA Words*, in Proceedings of the 32nd International Colloquium on Automata, Languages, and Programming (ICALP), Lisbon, Portugal, 2005, pp. 1275–1286.
- [3] B. PARDO AND M. SANGHI, *Polyphonic Musical Sequence Alignment for Database Search*, in Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR), London, England, 2005.
- [4] M. SANGHI, Z. LI, B. CHAVEZ, Y. CHEN, AND M.-Y. KAO, *Hamsa: Fast Signature Generation for Zero-day Polymorphic Worms with Provable Attack Resilience*, under submission.
- [5] S. SEN AND N. GUPTA, *Distribution-sensitive Algorithms*, Nordic Journal of Computing, 6 (1999), pp. 194–211.