CSoI Tier I seed grant proposal:

A Fresh Look at Boolean Functions

CSoI Post-doc advisor: **Thomas Courtade**
courtade@stanford.edu (Stanford/Princeton University)

CSoI Post-doc advisor: **Pulkit Grover**
pulkit@stanford.edu (Stanford University & Carnegie Mellon University)

Student PI: **Madars Virza**
madars@mit.edu (MIT, advisor: Professor R. Rivest)

September 1, 2012

I. Background and problem statement

A. Background

Formally, a boolean function on n inputs is given by $b : \{0,1\}^n \to \{0,1\}$. In other words, b maps a string of n bits to a single bit. While boolean functions form the basis of modern computation, we are still very far from a complete understanding of these functions. As a first example, Riordan and Shannon proved in 1942 (long before digital computation became prevalent) that almost every boolean function on n inputs requires roughly $2^n/n$ gates (AND, OR, and NOT) to implement [1]. However, the proof was nonconstructive, and to the present day, we still have not been able to prove a super linear lower bound on complexity for any *explicitly* given function![1]

As a second example, consider the following simple problem setup due to Kumar [3]. Let X^n be *i.i.d.* Bern$(1/2)$, Z^n be *i.i.d.* Bern(α), and $Y^n = X^n \oplus Z^n$, where $\oplus$ denotes the XOR operation. Can we determine the value of

$$\max_{b:\{0,1\}^n \to \{0,1\}} I(b(X^n); Y^n) \ ? \tag{1}$$

In other words, what is the most significant bit of information that X^n provides about Y^n. Intuitively, the value of (1) should be equal to $1 - H(\alpha)$. This is achieved by simply setting $b(X^n) = X_1$, however establishing the converse result has proved to be strikingly difficult and remains unfinished. It seems that the difficulty lies in dealing with functions with a fixed range. In fact, if we optimize over functions $f : \{0,1\}^n \to \{0,1\}^{k_n}$, where $\lim_{n\to\infty} k_n = \infty$, it is possible to show that

$$\max_{f:\{0,1\}^n \to \{0,1\}^{k_n}} \frac{1}{k_n} I(f(X^n); Y^n) \to (1-2\alpha)^2.$$

The key point is that nearly all information theoretic techniques rely on concentration of measure phenomena and non-asymptotic problem formulations are therefore much more difficult to deal with. Despite much recent work on non-asymptotic information theory, the fact remains that apparently simple problem formulations like (1) evade solution and therefore could point the way to new techniques and tools that should be developed.

Coincidentally, problems such as (1) have attracted the interest of computational biologists (cf. [4] and the references therein). This is primarily due to the fact that many cell functions can be represented as a boolean function; i.e., given the present conditions in a cell, is a particular protein produced?

B. Problem Statement

At this point, we have no precise problem formulation. Instead, our goal is to bring together researchers from the fields of information theory (Tom, Pulkit) and theoretical computer science (Madars) in order to take a fresh look at boolean functions from an interdisciplinary perspective. Each of us feels that there are interesting problems at the intersection of these fields which are of mutual interest to both communities. Several potential research directions we have already discussed include the following:

1) Towards one-way functions: Can we prove the existence of a function which requires a circuit complexity of n gates to compute and $m \gg n$ gates to invert? What changes if we require circuits to be efficiently 3D-embeddable?
2) "Practical" circuit constraints: If we impose realistic constraints on boolean circuits (such as being implementable in three dimensions), is it possible to derive stronger bounds that those for the traditional computational models?
3) Boolean function sensitivity: When considering problems such as (1), how does function sensitivity relate to mutual information? This is closely related to [5] and the computational biology works it cites.

II. Proposed Activity

During the award period for this Tier I proposal, our main goals are to develop a precise formulation of a research problem (leveraging the interdisciplinary composition of the team) and to develop a Tier II proposal based on this formulation and corresponding preliminary results. The problem we have described in Section I is related

[1]This is reminiscent of the folk theorem [2] "Almost all codes are good. Except those we know of." which held until the recent discovery of capacity achieving codes.

to the "Biological" thrust of the center, as boolean functions are of significant interest to researchers in the field of computational biology.

Successful formulation of a research problem and the subsequent development of preliminary results will require frequent interaction amongst the team members. Pulkit and Tom are both located at Stanford University until January, 2013 (at which time Pulkit will join Carnegie Mellon University), and therefore will permit biweekly meetings. Virtual meetings which include all group members will be held monthly. Moreover, the entire group will meet in person a minimum of two times. The virtual meetings will be held using the collaboration tools available through `soihub.org`.

The primary goal of these frequent meetings is to allow the group to work toward the expected outcomes, which are detailed in the following section.

III. Expected Outcomes

Since this is a Tier I proposal, we do not expect to answer the questions posed in Section I in their entirety during the award period. Rather, our goal is to bring together researchers with expertise in the relevant areas in order to formulate a precise research direction for which a Tier II proposal will be written. Specific outcomes include:

- **Short Term Outcomes**
  1) Thorough understanding of related work: Currently, each member of the proposed team is familiar with his/her specific area. However, one of the short term objectives is to combine this knowledge and also perform an in-depth literature survey to understand how the proposed problem fits into the context of existing work.
  2) Multi-institution collaboration via `soihub.org` tools: In order to facilitate multi-institution collaboration and track progress, we will use the collaboration tools available through `soihub.org`. Not only is this a required outcome for successfully funded projects, it will also facilitate project documentation and tracking.
  3) Precise problem formulation: As mentioned previously, one of the primary goals of this project is to exploit the interdisciplinary nature of the team in order to develop a clearly defined research problem related to the center's mission of promoting interdisciplinary research. Specifically, the research problem we formulate should be of mutual interest to several communities (information theory, computer science, and computational biology).

- **Long Term Outcomes**
  1) Tier II proposal: The ultimate goal of the work conducted during the award period for this Tier I proposal is to put together a set of well-formulated research problems which can be realistically solved. These problems will form the basis of a Tier II proposal which will be submitted next year.
  2) Conference paper or technical report: During the award period of this Tier I proposal, we anticipate that preliminary results will be established as we review the literature and work toward a precise formulation of the problem at hand. In particular, we hope to obtain partial results for at least one of the specific problems stated in Section I-B. If it is not possible to produce a conference paper based on these preliminary results, we will compile our findings into a technical report to support the Tier II proposal.

IV. Proposed Work Statement

We hope to build on expertise in three different areas — Tom's in information-theoretic limits on function computation, Pulkit's in information-theoretic limits in circuits, and Madars's in cryptography — to understand fundamental issues in boolean functions and one-way functions. To begin with, we will survey the related work in these fields, each of us taking the lead in the literature closest to our expertise. We will then discuss these papers in group meetings (virtual as well as in person) and follow-up with possible problem formulations. The goal is to arrive at a thorough understanding of the questions raised in Section I-B, and formulating and address precise, concrete versions of these questions based on this understanding.

As described above, we intend to have bi-weekly meetings in person for Pulkit and Tom, and at least monthly virtual meetings (using soihub tools). We also intend to meet in person at least two times.

References

[1] J. Riordan and C. E. Shannon, "The number of two-terminal series-parallel networks," J. Math. and Physics, no. 21, pp. 155–171, 1942.
[2] J. M. Wozencraft and B. Reiffen, "Sequential decoding," 1961.
[3] G. R. Kumar, "Csoi summer school poster," 2012.
[4] A. Samal and S. Jain, "The regulatory network of e. coli metabolism as a boolean dynamical system exhibits both homeostasis and flexibility of response," BMC Systems Biology, vol. 2, no. 1, p. 21, 2008. [Online]. Available: http://www.biomedcentral.com/1752-0509/2/21
[5] J. G. Klotz, D. Kracht, M. Bossert, and S. Schober, "Canalizing Boolean Functions Maximize the Mutual Information," ArXiv e-prints, Jul. 2012.
[6] T. A. Courtade, "Two Problems in Multiterminal Information Theory," Ph.D. dissertation, University of California, Los Angeles, 2012.

V. Budget and Justification

Budget section removed.

VI. Research statements

A. Thomas Courtade

Much of Tom's recent work has focused on lossy compression when the reproduction fidelity is measured under logarithmic loss. In this setting, the decompressor produces beliefs (i.e., soft decisions), rather than deterministic decision values. The logarithmic loss function is a method of measuring the quality of these beliefs with respect to the true realization of the data. Surprisingly, many longstanding open problems become tractable when studied in the setting where distortion is measured under logarithmic loss. These include the CEO problem, the multiterminal source coding problem, and the interactive lossy source coding problem, among others [6]. Hopefully, some of these newly developed tools can be brought to bear on the problems we propose to study.

Tom will serve as a center postdoc advisor for this project.

B. Pulkit Grover

My recent work deals with understanding circuit-complexity for circuits in a 3-D world on encoding and decoding in wireless communications. The traditional computational models — such as the Turing machine model and the traditional models for circuit complexity — do not constrain the communication to be performed in three dimensions. Simultaneously, obtaining lower bounds on Turing or circuit complexity of problems has proven to be extremely difficult. This is partly the reason that fundamental complexity lower bounds did not exist on encoding and decoding for coding techniques in information theory. My work derives fundamental lower bounds on wiring complexity of encoding and decoding error correcting codes by making explicit assumptions on 3-D modeling of computation. We're hoping some of those techniques can extend to lower bounds on other specific functions as well. I personally benefit immensely from this study because I am at a stage in my career where I am looking for new areas of research. Thorough this project, I get to learn about intellectual issues in cryptography and complexity, broadening my understanding ofi nformation security.

C. Madars Virza

I have previously worked on improving separation between two complexity measures of Boolean functions: sensitivity and block sensitivity. If proved to be superpolynomial, this separation would have great consequences for our understanding of complexity theory and provide new proof techniques for bounding, for example, randomized and quantum query complexities. My 2011 work presented improvement of the lower bound, since Rubinstein's function of 1995, however the general problem remains open: the best lower bound remains quadratic, while best upper bound is exponential. My recent work has been focused on cryptography (building public-key infrastructure resilient to key compromises) and I think that our research could greatly benefit the cryptography community. While circuit lower bounds have been extremely evasive (the best lower bound for explicit function is just linear), the cryptography community would be *very* happy to have lower bounds in more restrictive models, which accurately capture the physical reality, i.e. for circuits with efficient 3D embeddings. I come from theoretical computer science background and I am very looking forward to learning from information theorists of this proposal.