Enumeration and stratifiaction of Boolean functions by canalizing depth

Qijun He

Department of Mathematical Sciences
Clemson University
Clemson, SC

Network Dynamics and Simulation Science Laboratory (NDSSL)
Virginia Bioinformatics Institute (VBI)
Virginia Tech
Blacksburg, VA

July 15, 2015

My talk in one slide

Background

A Boolean function is canalizing if some variable can determine the output.

If that variable doesn't take its "canalizing input", the output is a function on $n-1$ variables:

$$g(\hat{x}_i) = g(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)\,.$$

We ask if it too is canalizing, and so on.

The nested canalizing functions (NCFs) are those that are "fully recursively canalizing."

NCFs are well-understood. They decompose nicely into "extended monomial layers" and have been enumerated. (Murrugarra et al., 2013)

Our contribution

Every Boolean function has a well-defined canalizing depth, k.

This allows us to decompose every Boolean function into extended monomial layers and a core polynomial.

This extends work by Murrugarra et al. on the *algebraic structure* of NCFs.

We derive enumeration formulas for "k-canalizing functions," which generalize known enumeration results for both canalizing functions and NCFs.

Canalizing & nested canalizing functions

Definition

A Boolean function $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ is canalizing if there exists a variable x_i, a Boolean function $g(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$, and $a, b \in \mathbb{F}_2$ such that

$$f(x_1, \dots, x_n) = \begin{cases} b & x_i = a, \\ g \not\equiv b & x_i \neq a. \end{cases}$$

In this case, x_i is a canalizing variable, the input a is the canalizing input, and the output value b when $x_i = a$ is the corresponding canalized output.

Definition

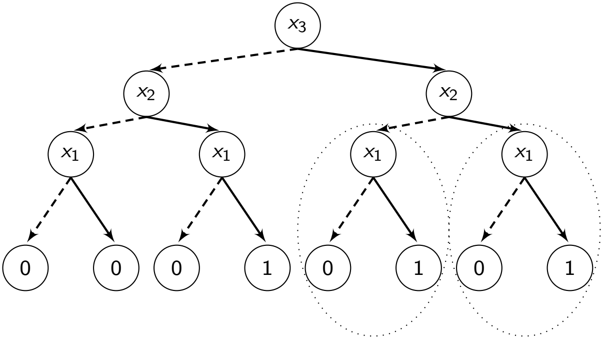
A function $f\colon \mathbb{F}_2^n \to \mathbb{F}_2$ is nested canalizing w.r.t. $\sigma \in \mathfrak{S}_n$, inputs a_i and outputs b_i, for $i = 1, 2, \dots, n$, if it can be represented as:

$$f(x_1, \dots, x_n) = \begin{cases} b_1 & x_{\sigma(1)} = a_1, \\ b_2 & x_{\sigma(1)} \neq a_1, x_{\sigma(2)} = a_2, \\ b_3 & x_{\sigma(1)} \neq a_1, x_{\sigma(2)} \neq a_2, x_{\sigma(3)} = a_3, \\ \vdots & \vdots \\ b_n & x_{\sigma(1)} \neq a_1, \dots, x_{\sigma(n-1)} \neq a_{n-1}, x_{\sigma(n)} = a_n, \\ \overline{b_n} & x_{\sigma(1)} \neq a_1, \dots, x_{\sigma(n-1)} \neq a_{n-1}, x_{\sigma(n)} \neq a_n. \end{cases}$$

Binary decision tree

A Boolean function can be evaluated using a binary decision tree and a fixed variable order.

For example, consider $f(x_1, x_2, x_3) = x_1x_2x_3 + x_1x_2 + x_1x_3$, with respect to variable order $x_3 < x_2 < x_1$.



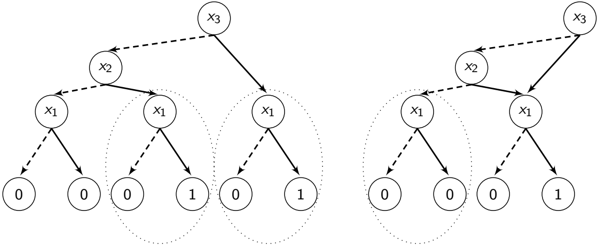
Binary decision diagram (BDD)

Reduction rules

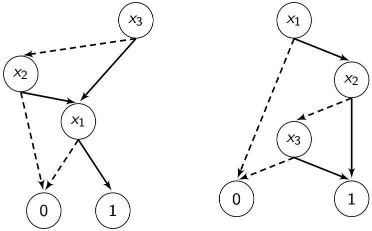
We can reduce the binary decision tree to a BDD by applying the following rules:

(i) Merge identical substructures that have the same parent node, and then eliminate that node.

(ii) Merge identical substructures that have different parents.



Average path length (APL) of BDDs



$$APL_f^{x_3<x_2<x_1} = (2 \cdot 6 + 3 \cdot 2)/8 = \frac{9}{4}$$

$$APL_f^{x_1<x_2<x_3} = (1 \cdot 4 + 2 \cdot 2 + 3 \cdot 2)/8 = \frac{7}{4}$$

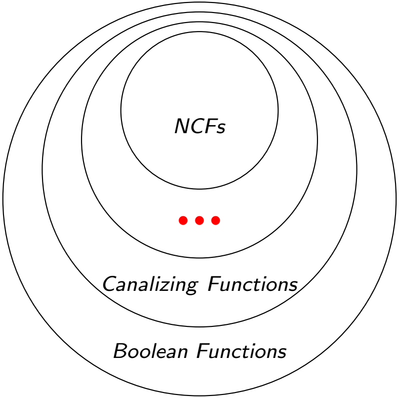
Theorem (Jarrah et al., 2007)

The n-variable Boolean functions with no ficticious variables that have minimum APL $(2 - \frac{1}{2^{n-1}})$ are precisely the nested canalizing functions.

Remark

The n-variable parity function (and its negation) has maximum APL of n.

A cartoon of all 2^{2^n} Boolean functions on n variables.



Not quite nested canalizing

Many functions are canalizing, but not nested canalizing.

For example, in a Boolean model of the *lac* operon by Robeva/Davies (2013):

$$f_L(t+1) = \overline{G_e} \wedge [(L \wedge \overline{E}) \vee (L_e \wedge E)] .$$

This means "*internal lactose (L) will be present the following timestep if there is no external glucose (G_e), and at least one of the following holds*":

- internal lactose is present, but the enzyme β-galactosidase (E) that breaks it down is absent;
- external lactose (L_e) is available *and* the *lac* permease transporter protein (E) is present.

Note: $\overline{G_e}$ is canalizing because it acts as a "shut-down" switch: if $G_e = 1$, then $f_L = 0$ regardless of the other variables. Thus,

$$f_L(G_e, L_e, L, E) = \begin{cases} 0 & G_e = 1, \\ (L \wedge \overline{E}) \vee (L_e \wedge E) & G_e \neq 0. \end{cases}$$

The function $g = (L_e \wedge E) \vee (L \wedge \overline{E})$ is not canalizing, and so f_L is canalizing but not nested canalizing.

k-canalizing functions

Definition

A Boolean function $f(x_1, \ldots, x_n)$ is k-canalizing, where $0 \leq k \leq n$, w.r.t. $\sigma \in \mathfrak{S}_n$, inputs a_i, and outputs b_i, for $1 \leq i \leq k$, if

$$f(x_1, \ldots, x_n) = \begin{cases} b_1 & x_{\sigma(1)} = a_1, \\ b_2 & x_{\sigma(1)} \neq a_1, x_{\sigma(2)} = a_2, \\ b_3 & x_{\sigma(1)} \neq a_1, x_{\sigma(2)} \neq a_2, x_{\sigma(3)} = a_3, \\ \vdots & \vdots \\ b_k & x_{\sigma(1)} \neq a_1, \ldots, x_{\sigma(k-1)} \neq a_{k-1}, x_{\sigma(k)} = a_k, \\ g \not\equiv b_k & x_{\sigma(1)} \neq a_1, \ldots, x_{\sigma(k-1)} \neq a_{k-1}, x_{\sigma(k)} \neq a_k. \end{cases}$$

where $g = g(x_{\sigma(k+1)}, \ldots, x_{\sigma(n)})$. When g is non-canalizing, k is the canalizing depth of f. If g is non-constant, it is the core function of f, denoted f_C.

Remark

Since $g \not\equiv b_k$, a function f that is k-canalizing with respect to $\sigma \in \mathfrak{S}_n$, inputs a_i and outputs b_i is essential in each $x_{\sigma(i)}$ for $i = 1, \ldots, k$.

k-canalizing functions

Example

The Boolean function $f(x, y, z, w) = xy(z + w)$ has canalizing depth 2 and core function $f_C = z + w$.

$$f(x,y,z,w) = \begin{cases} 0 & x = 0 \\ 0 & x \neq 0,\ y = 0 \\ z + w & x \neq 0,\ y \neq 0 \end{cases}$$

Remarks

In our framework, if we consider the set of all Boolean functions on n variables, then:

- The canalizing depth of a k-canalizing function is at least k.
- A non-canalizing function has canalizing depth 0, and if it is non-constant, then $f_C = f$.
- Every Boolean function is 0-canalizing.
- 1-canalizing means "canalizing."
- n-canalizing means "nested canalizing."
- If f has canalizing depth k and g is constant, then f has $n - k$ fictitious variables, and is an NCF its k essential variables.

Polynomial form

The set of Boolean functions on n variables is isomorphic to the quotient ring

$$R := \mathbb{F}_2[x_1, \ldots, x_n]/I, \qquad \text{where} \quad I = \langle x_i^2 - x_i : 1 \leq i \leq n \rangle .$$

Thus, a Boolean function f can be uniquely expressed as a square-free polynomial – its "algebraic normal form."

Lemma (Murrugarra et al., 2013)

$f(x_1, \ldots, x_n)$ is canalizing in x_i, with input a_i and output b_i, iff for some polynomial $g \not\equiv 0$,

$$f = (x_i + a_i)g(\hat{x}_i) + b_i \, .$$

Theorem

$f(x_1, \ldots, x_n)$ is k-canalizing, w.r.t. $\sigma \in \mathfrak{S}_n$, inputs a_i and outputs b_i, for $1 \leq i \leq k$, iff it has polynomial form

$$f(x_1, \ldots, x_n) = (x_{\sigma(1)} + a_1)g(\hat{x}_i) + b_1 \, ,$$

where

$$g(\hat{x}_i) = (x_{\sigma(2)} + a_2)\Big[\ldots \big[(x_{\sigma(k-1)} + a_{k-1})[(x_{\sigma(k)} + a_k)\bar{g} + \Delta b_{k-1}] + \Delta b_{k-2} \big] \ldots \Big] + \Delta b_1$$

for some polynomial $\bar{g} = \bar{g}(x_{\sigma(k+1)}, \ldots, x_{\sigma(n)}) \not\equiv 0$, where $\Delta b_i := b_{i+1} - b_i = b_{i+1} + b_i$.

Extended monomial layers

Definition

A Boolean function $M(x_1, \dots, x_m)$ is an extended monomial in variables $x_1, \dots, x_m$ if

$$M(x_1, \dots, x_m) = \prod_{i=1}^{m} (x_i + a_i),$$

where $a_i \in \mathbb{F}_2$ for each $i = 1, \dots, m$.

(Murrugarra et al., 2013)

A function $f(x_1, \dots, x_n)$ is an NCF iff

$$f(x_1, \dots, x_n) = M_1(M_2(\cdots(M_{r-1}(M_r + 1) + 1)\cdots) + 1) + b$$

for extended monomials M_i with disjoint supports, and $b \in \mathbb{F}_2$.

Special case of this: all variables are canalizing iff $f = M(x_1, \dots, x_n) + b$, where M is an extended monomial in all variables.

Extended monomial layers & core polynomials

Not only can NCFs be written in disjoint extended monomial layers, but so can all Boolean functions.

Theorem

Every Boolean function $f(x_1, \ldots, x_n) \not\equiv 0$ can be uniquely written as

$$f(x_1, \ldots, x_n) = M_1(M_2(\cdots(M_{r-1}(M_r p_C + 1) + 1)\cdots) + 1) + b, \tag{1}$$

where each $M_i = \prod_{j=1}^{k_i}(x_{i_j} + a_{i_j})$ is a nonconstant extended monomial, $p_C \not\equiv 0$ is the core polynomial of f, and $k = \sum k_i$ is the canalizing depth. Each x_i appears in exactly one of $\{M_1, \ldots, M_r, p_C\}$, and the only restrictions on Eq. (1) are the following "exceptional cases":

(i) If $p_C \equiv 1$ and $r \neq 1$, then $k_r \geq 2$;

(ii) If $p_C \equiv 1$ and $r = 1$ and $k_1 = 1$, then $b = 0$;

When f is non-canalizing, then $p_C = f$.

Example

The Boolean function $f(x_1, \ldots, x_7) = x_1\overline{x_2}(x_3x_4(x_5 + x_6 + x_7 + 1) + 1)$ has canalizing depth 4. With respect to the permutation $\sigma = 1, 2, 3, 4$, its canalizing inputs are $(a_i)_{i=1}^4 = (0, 1, 0, 0)$, outputs $(b_i)_{i=1}^4 = (0, 0, 1, 1)$ and the core polynomial is $p_C = x_5 + x_6 + x_7 + 1$.

Extended monomial layers & core polynomials

Let's take a look at the last example a little more closely.

Example

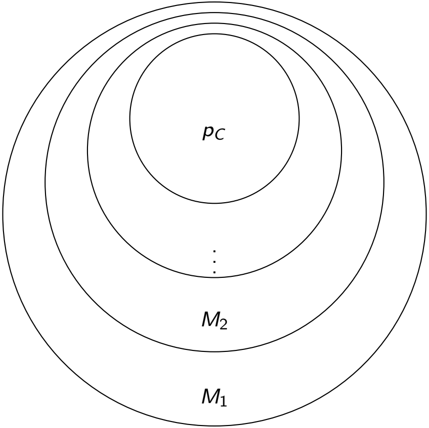
The Boolean function $f(x_1, \ldots, x_7) = x_1\overline{x_2}(x_3x_4(x_5 + x_6 + x_7 + 1) + 1)$ has canalizing depth 4. With respect to the permutation $\sigma = 1, 2, 3, 4$, its canalizing inputs are $(a_i)_{i=1}^4 = (0, 1, 0, 0)$, outputs $(b_i)_{i=1}^4 = (0, 0, 1, 1)$ and the core polynomial is $p_C = x_5 + x_6 + x_7 + 1$.

$$f(x_1, \ldots, x_7) = \begin{cases} 0 & x_1 = 0 \\ 0 & x_1 \neq 0, x_2 = 1 \\ 1 & x_1 \neq 0, x_2 \neq 1, x_3 = 0 \\ 1 & x_1 \neq 0, x_2 \neq 1, x_3 \neq 0, x_4 = 0 \\ x_5 + x_6 + x_7 & x_1 \neq 0, x_2 \neq 1, x_3 \neq 0, x_4 \neq 0 \end{cases}$$

There are 2^{2^n} Boolean functions on n variables. Each one has a unique well-defined:

- canalizing depth,
- extended monomial layers structure,
- core polynomial.

A cartoon of an arbitrary Boolean function



Enumeration

Prior results

- *Just/Shmulevich/Konvalina, 2004*: The number C_n of canalizing Boolean functions on $n \geq 0$ variables is

$$C_n = 2((-1)^n - n - 1) + \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} 2^{k+1} 2^{2^{n-k}}.$$

- *Murrugarra et al., 2013*: The number of NCFs on n variables is:

$$B(n,n) = 2^{n+1} \sum_{r=1}^{n-1} \sum_{\substack{k_1+\ldots+k_r=n \\ k_i \geq 1,\ k_r \geq 2}} \binom{n}{k_1, \ldots, k_r},$$

where $\binom{n}{k_1,\ldots,k_r} = \frac{n!}{k_1!k_2!\ldots k_r!}$.

Theorem

The number of Boolean functions on n variables with canalizing depth k is

$$B(n,k) = \binom{n}{k} \left[B(k,k) + B^*(n-k,0) \cdot 2^{k+1} \sum \binom{k}{k_1, \ldots, k_r} \right],$$

where the sum is taken over all *compositions of* k.

An example: $n = 4$

There are $2^{2^4} = 65536$ Boolean functions on 4 variables.

The number of functions with canalizing depth exactly k, for $k = 1, 2, 3, 4$ is

$$
\begin{aligned}
B(4,4) &= \binom{4}{4}(736 + 0) = 736 \\
B(4,3) &= \binom{4}{3}(64 + 0) = 256 \\
B(4,2) &= \binom{4}{2}(8 + 2 \cdot 8 \cdot 3) = 336. \\
B(4,1) &= \binom{4}{1}(2 + 136 \cdot 4 \cdot 1) = 2184.
\end{aligned}
$$

Summing these yields the total number of canalizing functions on 4 variables,

$$C_4 = 3512 = 736 + 256 + 336 + 2184 = B(4,4) + B(4,3) + B(4,2) + B(4,1).$$

Thus, there are $B(4,0) = 65536 - 3512 = 62024$ non-canalizing functions on four variables, including the two constant functions.

Average path length of k-canalizing functions

Recall that the APL an n-variable function is:

- minimized for the NCFs (i.e., n-canalizing functions); APL$= 2 - \frac{1}{2^{n-1}}$.
- maximized for Par and 1+Par (which are 0-canalizing); APL$= n$.

Theorem (H.)

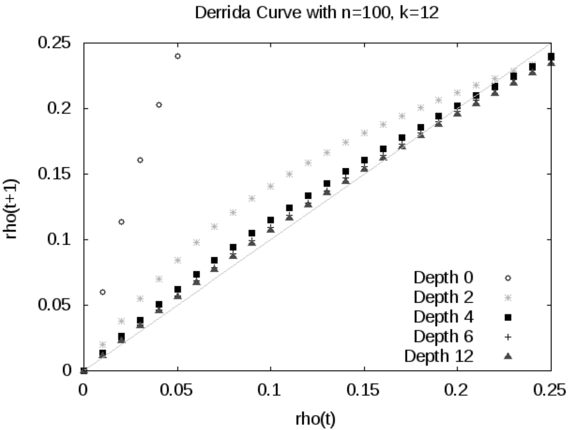
If f is k-canalizing ($k \geq 1$) on n variables, then the APL of its BDD is:

$$2 - \frac{1}{2^{k-1}} \leq APL_f \leq 2 - \frac{1}{2^{k-1}} + \frac{n-k}{2^k}$$

Network dynamics of k-canalizing functions

Graph dynamical systems built with k-canalizing functions are more stable than random networks.

Moreover, the stability increases with k. This can be measured using a Derrida plot.



Current and future research

In the future, we plan to:

- derive asymptotics for the number of n-variables Boolean functions of canalizing depth k, as n and k grow large;
- investigate well-known Boolean network models and compute the canalizing depth of the proposed functions;
- design reverse-engineering algorithms for Boolean networks models from partial data using k-canalizing functions;
- investigate whether the set of k-canalizing functions that fit the model space has an inherent algebraic structure (e.g., toric variety?);
- extend the results from this talk from Boolean to multi-state functions.

Open-ended questions for NDSSL: In GDS theory, many computationally hard algorithms become tractable for special classes of functions and/or graphs, such as

- bounded tree width
- k-symmetric functions

I'm particularly interested in discussing with you whether anything can be said about these problems for k-canalizing functions.

References

[1] R. Albert and H.G. Othmer. **The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in drosophila melanogaster**. *J. Theor. Biol.*, 223(1):1–18, 2003.

[2] Q. He and M. Macauley. **Enumeration and stratification of Boolean functions by canalizing depth**. Submitted to *Physica D*, 2015.

[3] F. Hinkelmann and A.S. Jarrah. **Inferring biologically relevant models: nested canalyzing functions**. *ISRN Biomathematics*, 2012.

[4] A.S. Jarrah and R. Laubenbacher. **Discrete models of biochemical networks: The toric variety of nested canalyzing functions**. In *Algebraic Biology*, pages 15–22. Springer, 2007.

[5] A.S. Jarrah, B. Raposa, and R. Laubenbacher. **Nested canalyzing, unate cascade, and polynomial functions**. *Physica D*, 233(2):167–174, 2007.

[6] W. Just, I. Shmulevich, and J. Konvalina. **The number and probability of canalizing functions**. *Physica D*, 197(3):211–221, 2004.

[7] C. Kadelka, Y. Li, J.O. Adeyeye, and R. Laubenbacher. **Nested canalizing functions and their networks**. *arXiv:1411.4067*, 2014.

[8] Y. Li, J.O. Adeyeye, D. Murrugarra, B. Aguilar, and R. Laubenbacher. **Boolean nested canalizing functions: A comprehensive analysis**. *Theor. Comput. Sci.*, 481:24–36, 2013.

[9] L. Layne, E. Dimitrova, and M. Macauley. **Nested canalyzing depth and network stability**. *Bull. Math. Biol.*, 74(2):422–433, 2012.

[10] D. Murrugarra and R. Laubenbacher. **The number of multistate nested canalyzing functions**. *Physica D*, 241(10):929–938, 2012.

[11] R. Robeva and T. Hodge. **Mathematical concepts and methods in modern biology: using modern discrete models**. Academic Press, 2013.

[12] . Veliz-Cuba and B. Stigler. **Boolean models can explain bistability in the lac operon**. *J. Computat. Biol.*, 18(6):783–794, 2011.