The Kauffman $N^K$ Model
A Stochastic Combinatorial Optimization Model for Complex Systems

Hemanshu Kaul
hkaul@math.uiuc.edu.

University of Illinois at Urbana-Champaign
We want to model systems composed of several interacting components, where each component can be in one of many possible states.

Objective: Maximize a measure of performance of the system based on contributions from each component, depending on the state of the component and its ‘interaction’ with its neighbors.
In 1987, Kauffman and Levin introduced the NK model.

- $N$ counts the number of components in the system.
- $K$ measures the ‘degree’ of interaction between components.

The NK model was originally proposed to study the evolution of genomes.

- system $\equiv$ genome
- components $\equiv$ genes
- states $\equiv$ gene mutations
- performance measure $\equiv$ fitness
Mathematical Description

System – A vector with $N$ components, each of which can be in one of 2 possible states. $\mathbf{x} = (x_0, \ldots, x_{N-1})$, with $x_i \in \{0, 1\}$ and the numbers 0, 1 used as labels for the states.

Performance Measure –

$$\Phi(\mathbf{x}) = \frac{1}{N} \sum_{i=0}^{N-1} \phi_i(\mathbf{x})$$

$\phi_i(\mathbf{x})$ is the performance contribution from each component $i$. 
Performance Measure –

\[ \Phi(x) = \frac{1}{N} \sum_{i=0}^{N-1} \phi_i(x) \]

\( \phi_i \), the contribution of component \( i \) to the overall performance of the system depends on

- its own state, and
- the states of \( K \) ‘neighboring’ components.
Performance Measure

\[ N = 6 \text{ and } K = 3 \]

System \((0, 1, 1, 0, 1, 0)\)

\[
\begin{array}{cccccc}
0 & 1 & 1 & 0 & 1 & 0 \\
\end{array}
\]

\[
\begin{array}{cc}
0 & 1 \\
1 & 1 \\
0 & \\
\end{array}
\]
\[ \Phi(x) = \frac{1}{N} \sum_{i=0}^{N-1} \phi_i(x_i, \ldots, x_{i+K}) \]

where arithmetic in the subscripts is done modulo \( N \) and \( \phi_i \) are \( N \) distinct real-valued functions on \( \{0, 1\}^{K+1} \).

\( N = 4 \) and \( K = 2 \)

\[ \Phi(0, 1, 1, 0) = \frac{1}{4}[\phi_0(0, 1, 1)+\phi_1(1, 1, 0)+\phi_2(1, 0, 0)+\phi_3(0, 0, 1)] \]
Question – Given \( N, K, 0 \leq K \leq N - 1 \), and \( \phi_i : \{0, 1\}^{K+1} \rightarrow \mathbb{R} \), \( i = 0, 1, \ldots, N - 1 \)

How can we find a system with the best possible performance?

\[
\max\{\Phi(x) \mid x \in \{0, 1\}^N\}
\]
$N = 4$ and $K = 2$
$2^N = 2^4 = 16$ possible systems

\[ \Phi(0, 0, 0, 0) = \frac{1}{4} [\phi_0(0, 0, 0) + \phi_1(0, 0, 0) + \phi_2(0, 0, 0) + \phi_3(0, 0, 0)] \]

\[ \Phi(0, 0, 1, 0) = \frac{1}{4} [\phi_0(0, 0, 1) + \phi_1(0, 1, 0) + \phi_2(1, 0, 0) + \phi_3(0, 0, 0)] \]

\[ \Phi(0, 1, 1, 0) = \frac{1}{4} [\phi_0(0, 1, 1) + \phi_1(1, 1, 0) + \phi_2(1, 0, 0) + \phi_3(0, 0, 1)] \]

\[ \Phi(0, 1, 1, 1) = \frac{1}{4} [\phi_0(0, 1, 1) + \phi_1(1, 1, 1) + \phi_2(1, 1, 0) + \phi_3(1, 0, 1)] \]

\[ \Phi(1, 1, 1, 1) = \frac{1}{4} [\phi_0(1, 1, 1) + \phi_1(1, 1, 1) + \phi_2(1, 1, 1) + \phi_3(1, 1, 1)] \]
Central Question

Given $N$, $K$, $0 \leq K \leq N - 1$, and $\phi_i : \{0, 1\}^{K+1} \rightarrow \mathbb{R}$, $i = 0, 1, \ldots, N - 1$

What can we say about the Global Optima, the system that maximizes the value of the performance measure?

$$\max \{ \Phi(x) \mid x \in \{0, 1\}^N \}$$

- NP-complete problem.
- In applications it is difficult, if not impossible, to determine the values taken by $\phi_i$.

So, this combinatorial optimization problem is formulated and studied probabilistically.
Given $N$, $K$, with $0 \leq K \leq N - 1$, and $N2^{K+1}$ random variables $\phi_i(y)$ for $y \in \{0, 1\}^{K+1}$, $i = 0, 1, \ldots, N-1$, independently and identically distributed as $F$.

Study the distribution of the global optima –

$$X_{N,K} = \max\{\Phi(x) \mid x \in \{0, 1\}^N\}$$

where $\Phi(x) = \frac{1}{N} \sum_{i=0}^{N-1} \phi_i(x_i, \ldots, x_i+K)$. 
Research Question– How do the varying values of $N$ and $K$ affect the performance of the systems?
Research Question– How do the varying values of $N$ and $K$ affect the performance of the systems?

- Mostly study of local optima w.r.t. a Hamming distance based neighborhood structure.
- Mostly simulation-based results and applications.
- Solow et. al (2000) showed the global decision problem is NP-complete.
Previous Research

Evans and Steinsaltz (2002)
- convert to an infinite-dimensional variational problem
- explicit bounds only when $K = 1$ and $F$ is exponential distribution

Durrett and Limic (2003)
- use the theory of substochastic Harris chains
- explicit bounds only when $K = 1$ and $F$ is negative exponential distribution

Numerous other papers (both Applications and Theory).
Summary of Results

- Develop a simple computational set-up, independent of the underlying distribution $F$.
- Develop methodology for finding bounds on the moments of $X_{N,K}$, independent of the underlying distribution $F$.
- Find explicit bounds on the expectation of $X_{N,K}$ when $K$ is function of $N$, for fundamental underlying distributions like uniform and normal.
- Show concentration of $X_{N,K}$ around its mean, $E_{N,K}$.

We use tools from Combinatorics and Graph Theory, Networks, Probability and Statistics, and Geometry.
Network $D_{N,K}$: 2^{K+1} \times (N + 1)$ array of vertices, $v^i_t$, $t \in \{0, 1\}^{K+1}$, 0 \leq i \leq N.

Each vertex, $v^i_t$, corresponds to component $i$ and $t$, the state vector for the component and its $K$ neighbors.

Idea – Create a correspondence between the systems and the directed paths in this network.

$v^i_t \rightarrow v^j_{\hat{t}_t} \iff j = i + 1$ and $\hat{t}_i = t_{i+1}$, $i = 1, \ldots, K$

and $\hat{t}_{K+1} \in \{0, 1\}$

Each $v^i_t$ has a weight generated by the performance contribution (and random variable) $\phi_i(t)$. 
Network $D_{N,K}$

$N = 4$ and $K = 1$

\[
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \\
4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 & 4 \\
\end{array}
\]

Green path corresponds to the system \(\{0, 0, 1, 0\}\) and the weight of the path is the performance measure of the system.

Each directed path from from $v_t^0$ to $v_t^N$

and its associated weight

\[\uparrow\]

Each system and its performance
\( D_{N,K}^t \equiv \text{subnetwork of } D_{N,K} \text{ defined by all the directed paths between } v_0^t \text{ and } v_N^t \)
Computational Strategy for $K$ close to $N$

Observation – The value of $N - K$ determines the general structure of subnetwork $D_{N,K}^t$, while $N$ determines its size.

Subnetwork $D_{N,K}^0$ for $N - K = 2$

This leads to a recursive scheme whose each recursive step reduces the value of $N$ and brings it closer to the (fixed) value of $K$, until $N = K + 1$, in $N - K$ steps.
\( D'_{N,K} \) – Computational Strategy for small \( K \)

\( D'_{N,K} \equiv \) Network formed from \( D_{N,K} \) by deleting the vertices in the \( K+1 \) columns from \( N - K \) to \( N \) and adding a source and a sink.

Each directed path in \( D'_{N,K} \) corresponds to a unique system, but not all feasible systems are represented by a path in \( D'_{N,K} \).

\[
X_{N,K} \geq \frac{1}{N} \left[ l'_{N,K} + \sum_{i=N-K}^{N-1} X_i \right], \quad X_i \text{ i.i.d. } F
\]

\( l'_{N,K} \equiv \) maximum weight of a directed path in \( D'_{N,K} \)
$D''_{N,K}$ – Computational Strategy for small $K$

$D''_{N,K} \equiv$ Network formed from $D_{N,K}$ by deleting the vertices in column $N$ and adding a source and a sink.

Each feasible system corresponds to a unique directed path in $D''_{N,K}$, but not all directed paths represent a system.

$X_{N,K} \leq \frac{1}{N} \left[ l''_{N,K} \right]$

$l''_{N,K} \equiv$ maximum weight of a directed path in $D''_{N,K}$
$D'_{N,K}$ and $D''_{N,K}$

$D'_{N,K}$ has $N - K$ columns and $D''_{N,K}$ has $N$ columns.

For fixed $K$, the bounds in terms of $i'_{N,K}$ and $i''_{N,K}$ will be asymptotically tight.
Dependence between $\Phi(x)$ and $\Phi(y)$, $x, y \in \{0, 1\}^N$

$\Phi(x) = \frac{1}{N} \sum_{i=0}^{N-1} \phi_i(x_i, \ldots, x_{i+K})$

$\Phi(y) = \frac{1}{N} \sum_{i=0}^{N-1} \phi_i(y_i, \ldots, y_{i+K})$

$\Phi(x)$ and $\Phi(y)$ are dependent $\iff$ there exists $i$ such that $x_j = y_j$ for $i \leq j \leq i + K$

$G_{N,K} \equiv$ dependency graph for given $N, K$

vertices $\equiv x \in \{0, 1\}^N$

$x \leftrightarrow y \iff \Phi(x)$ and $\Phi(y)$ are dependent
Want to partition the vertex set of $G_{N,K}$, $V(G_{N,K}) = V_1 \sqcup V_2 \sqcup \ldots \sqcup V_t$, such that

- there are no edges within each class $V_i$
- sizes of any two classes differ by at most 1

$t$-equitable coloring of $G_{N,K}$

Theorem: $G_{N,K}$ has a $t$-equitable coloring if $t > N2^{N-K-2}$
**Bounds on Order Statistics**

Notation: \( Y_{[n]} = \max \{Y_1, \ldots, Y_n\} \)

\[ F_N \equiv \text{distribution of } \sum_{i=1}^{N} X_i \text{, for } X_i \text{ i.i.d. } F \]

\( X_{N,K} = \frac{1}{N} Y_{[2^N]} \), \( Y_i \sim F_N \); \( \{Y_i\} = \{\Phi(x)\} \) dependent

**Theorem:** For all \( N, K \), with underlying distribution \( F \),

\[ \mathbb{E}[Y_{[2^{K+2}/N]}] \leq \mathbb{E}[X_{N,K}] \leq \mathbb{E}[Y_{[2^{K+2}/N]}] + \sqrt{N2^{N-K-2}\text{Var}[Y_{[2^{K+2}/N]}]} \]

where \( Y_1, \ldots, Y_k \text{ i.i.d. } F_N \).

*Proofs* use tools from Order Statistics and the Equitable Coloring of Graphs.
Bounds when $F = \mathbf{n}(0, 1)$

Theorem: For all $N \geq 2$, $K = N - 1$,

$$\sqrt{2 \log 2} - \frac{o(1)}{\sqrt{N}} \leq \mathbb{E}[X_{N,K}] \leq \sqrt{(1 + \frac{1}{N})2 \log 2} - \frac{o(1)}{\sqrt{N}}$$

Theorem: For all $N \geq 2$, $K = N - \alpha$, $\alpha \in \mathbb{Z}^+$, $\alpha \geq 2$, $c = \alpha - 2$

$$\sqrt{(1 - \frac{c}{N})2 \log 2 - \frac{2 \log N}{N}} - \frac{o(1)}{\sqrt{N}} \leq \mathbb{E}[X_{N,K}] \leq \sqrt{(1 + \frac{1}{N})2 \log 2} - \frac{o(1)}{\sqrt{N}}$$

Theorem: For all $N \geq 2$, $K = \beta N$, $\beta \in (0, 1)$

$$\sqrt{(\beta + \frac{2}{N})2 \log 2 - \frac{2 \log N}{N}} - \frac{o(1)}{\sqrt{N}} \leq \mathbb{E}[X_{N,K}] \leq \sqrt{(1 + \frac{1}{N})2 \log 2} - \frac{o(1)}{\sqrt{N}}$$

Leading Coefficients in both upper and lower bounds are equal to $\sqrt{2 \log 2}$
Bounds when $F = u(0, 1)$

When $\{X_j\}$ i.i.d. $u(0, 1)$,

$\Pr\{\sum_{j=1}^{N} X_j \leq x\}$ is equal to the volume of

$P(x) = \{y \in \mathbb{R}^N \mid \sum_{j=1}^{N} y_j \leq x \text{ and } 0 \leq y_j \leq 1\}$ a subset of

the $N$-dimensional hypercube $[0, 1]^N$.

We prove lemmas about $Vol(P(x))$ that help to decompose the expectation integral.

For Example,

Lemma : If $x > (1 - \frac{1}{2e})N$, then $Vol(P(x)) \geq 1 - \frac{1}{\sqrt{2\pi N} 2^N}$ for all $N \geq 2$. 
Bounds when \( F = \mathbf{u}(0, 1) \)

Theorem : For all \( N \geq 2 \), \( K = N - 1 \),

\[
(1 - \frac{1}{4}(2N)^{1/N}) \left(1 - \left(1 - \frac{N}{2N}\right)^{2N}\right) \leq \mathbb{E}[X_{N,K}] \leq 1 - \frac{1}{2e} \left(1 - \frac{1}{\sqrt{2\pi N} 2^N}\right)^{2N}
\]

\[
\lim_{N \to \infty} \text{Var}[X_{N,K}] \leq \frac{7}{16} - \frac{1}{e} \left(1 - \frac{1}{2e}\right) \approx 0.1373
\]

Theorem : For all \( N \geq 2 \), \( K = N - \alpha \), \( \alpha \in \mathbb{Z}^+ \), \( \alpha \geq 2 \), \( c = \alpha - 2 \)

\[
(1 - \frac{1}{4}(2N)^{1/N}) \left(1 - \left(1 - \frac{N}{2N}\right)^{\frac{2N}{cN}}\right) \leq \mathbb{E}[X_{N,K}] \leq 1 - \frac{1}{2e} \left(1 - \frac{1}{\sqrt{2\pi N} 2^N}\right)^{2N}
\]

Theorem : For all \( N \geq 2 \), \( K = \beta N \), \( \beta \in (0, 1) \)

\[
(1 - \frac{1}{4}(2N)^{1/N}) \left(1 - \left(1 - \frac{N}{2N}\right)^{4\frac{2\beta N}{N}}\right) \leq \mathbb{E}[X_{N,K}] \leq 1 - \frac{1}{2e} \left(1 - \frac{1}{\sqrt{2\pi N} 2^N}\right)^{2N}
\]

Leading Coefficients : \( 1 - \frac{1}{4} = 0.75 \) and \( 1 - \frac{1}{2e} \approx 0.816 \)
Probability of $X_{N,K}$ being far from $E[X_{N,K}]$ is exponentially decaying.

**Theorem:** If $F$ is a bounded distribution such that $X \sim F \Rightarrow |X| \leq c$, then

$$P[ | X_{N,K} - E[X_{N,K}] | \geq t ] \leq 2 \exp \left( -\frac{2Nt^2}{c^22^{2N-K-1}} \right)$$

**Proof** using Independent Bounded Differences Inequality, a variant of Azuma’s Martingale inequality.