Sports Leagues and Organization of Competitive Societies

Eli Ben-Naim
Los Alamos National Laboratory

Review Article
Randomness in Competitions
EB, N. Hengartner, S. Redner, and F. Vazquez

IST Young Scientists Symposium: Self-Organizing Systems, April 8, 2015

Talk, papers available from: http://cnls.lanl.gov/~ebn
What is the most competitive sport?

Football  
Baseball  
Hockey  
Basketball  
American Football
What is the most competitive sport?

- Football
- Baseball
- Hockey
- Basketball
- American Football

Can competitiveness be quantified?
How can competitiveness be quantified?
I. Modeling competitions
Parity of a sports league

- Teams ranked by win-loss record
- Win percentage:
  \[ x = \frac{\text{Number of wins}}{\text{Number of games}} \]
- Standard deviation in win-percentage:
  \[ \sigma = \sqrt{\langle x^2 \rangle - \langle x \rangle^2} \]
- Cumulative distribution = Fraction of teams with winning percentage < x
  \[ F(x) \]

Austrian Bundesliga

0.300 < x < 0.700
\[ \sigma = 0.13 \]
Parity by year

\[ \max = \frac{1}{2\sqrt{3}} = 0.288675 \]

\[ \langle \sigma \rangle = 0.13 \]

Parity is fluctuating, without clear trend
Data

- **300,000 Regular season games** (all games ever played)
- **5 Major** sports leagues in North America & England

<table>
<thead>
<tr>
<th>sport</th>
<th>league</th>
<th>full name</th>
<th>country</th>
<th>years</th>
<th>games</th>
</tr>
</thead>
<tbody>
<tr>
<td>football</td>
<td>FA</td>
<td>Football Association</td>
<td>🇬🇧</td>
<td>1888-2005</td>
<td>43,350</td>
</tr>
<tr>
<td>baseball</td>
<td>MLB</td>
<td>Major League Baseball</td>
<td>🇺🇸 🇨🇦</td>
<td>1901-2005</td>
<td>163,720</td>
</tr>
<tr>
<td>hockey</td>
<td>NHL</td>
<td>National Hockey League</td>
<td>🇨🇦</td>
<td>1917-2005</td>
<td>39,563</td>
</tr>
<tr>
<td>basketball</td>
<td>NBA</td>
<td>National Basketball Association</td>
<td>🇺🇸</td>
<td>1946-2005</td>
<td>43,254</td>
</tr>
<tr>
<td>american football</td>
<td>NFL</td>
<td>National Football League</td>
<td>🇺🇸</td>
<td>1922-2004</td>
<td>11,770</td>
</tr>
</tbody>
</table>

Standard deviation in winning percentage

Distribution of winning percentage clearly distinguishes sports

• Baseball most competitive?
• American football least competitive?

Fort and Quirk, 1995
The competition model

- Two, randomly selected, teams play
- Outcome of game depends on team record
  - Weaker team wins with probability \( q < 1/2 \)
  - Stronger team wins with probability \( p > 1/2 \)
  - When two equal teams play, winner picked randomly
- Initially, all teams are equal (0 wins, 0 losses)
- Teams play once per unit time

\[
\begin{align*}
(i, j) & \rightarrow \begin{cases} 
(i + 1, j) & \text{probability } p \\
(i, j + 1) & \text{probability } 1 - p
\end{cases} \\
q & = 1/2 \\
q & = 0 \quad \text{random} \\
q & = 0 \quad \text{deterministic} \\
p + q & = 1
\end{align*}
\]
Rate equation approach

- **Probability distribution functions**
  \( g_k = \text{fraction of teams with} \ k \ \text{wins} \)
  \( G_k = \sum_{j=0}^{k-1} g_j = \text{fraction of teams with less than} \ k \ \text{wins} \)
  \( H_k = 1 - G_{k+1} = \sum_{j=k+1}^{\infty} g_j \)

- **Evolution of the probability distribution**
  \[
  \frac{dg_k}{dt} = (1 - q)(g_{k-1}G_{k-1} - g_kG_k) + q(g_{k-1}H_{k-1} - g_kH_k) + \frac{1}{2} (g_{k-1}^2 - g_k^2)
  \]
  better team wins \quad worse \text{ team wins} \quad equal \text{ teams play}

- **Closed equations for the cumulative distribution**
  \[
  \frac{dG_k}{dt} = q(G_{k-1} - G_k) + (1/2 - q) (G_{k-1}^2 - G_k^2)
  \]

  **Boundary Conditions** \( G_0 = 0 \quad G_\infty = 1 \)
  **Initial Conditions** \( G_k(t = 0) = 1 \)

**Nonlinear Difference-Differential Equations**

*A kinetic view of statistical physics*, Kravivsky, Redner, EB, Cambridge University Press, 2010
An exact solution

- Stronger always wins ($q=0$)
  \[
  \frac{dG_k}{dt} = G_k(G_k - G_{k-1})
  \]

- Transformation into a ratio
  \[
  G_k = \frac{P_k}{P_{k+1}}
  \]

- Nonlinear equations reduce to linear recursion
  \[
  \frac{dP_k}{dt} = P_{k-1}
  \]

- Integrable (discrete) Burgers equation!
  \[
  G_k = \frac{1 + t + \frac{1}{2!} t^2 + \cdots + \frac{1}{k!} t^k}{1 + t + \frac{1}{2!} t^2 + \cdots + \frac{1}{(k+1)!} t^{k+1}}
  \]

EB, Krapivsky
J Phys A 2012
Long-time asymptotics

- Long-time limit
  \[ G_k \to \frac{k + 1}{t} \]

- Scaling form
  \[ G_k \to F \left( \frac{k}{t} \right) \]

- Scaling function
  \[ F(x) = x \]

Seek similarity solutions
Use winning percentage as scaling variable
Scaling analysis

- Rate equation
  \[
  \frac{dG_k}{dt} = q(G_{k-1} - G_k) + \left(\frac{1}{2} - q\right)(G_{k-1}^2 - G_k^2)
  \]

- Treat number of wins as continuous
  \[
  G_{k+1} - G_k \to \frac{\partial G}{\partial k}
  \]

- Inviscid Burgers equation
  \[
  \frac{\partial v}{\partial t} + v \frac{\partial v}{\partial x} = 0
  \]

- Stationary distribution of winning percentage
  \[
  G_k(t) \to F(x) \quad x = \frac{k}{t}
  \]

- Scaling equation
  \[
  \left[(x - q) - (1 - 2q)F(x)\right] \frac{dF}{dx} = 0
  \]
Scaling solution

• Stationary distribution of winning percentage

\[ F(x) = \begin{cases} 
0 & 0 < x < q \\
\frac{x - q}{1 - 2q} & q < x < 1 - q \\
1 & 1 - q < x. 
\end{cases} \]

• Distribution of winning percentage is uniform

\[ f(x) = F'(x) = \begin{cases} 
0 & 0 < x < q \\
\frac{1}{1 - 2q} & q < x < 1 - q \\
0 & 1 - q < x. 
\end{cases} \]

• Variance in winning percentage

\[ \sigma = \frac{1/2 - q}{\sqrt{3}} \quad \rightarrow \begin{cases} 
q = 1/2 & \text{perfect parity} \\
q = 0 & \text{maximum disparity} 
\end{cases} \]
Approach to scaling

Numerical integration of the rate equations, $q=1/4$

- Winning percentage distribution approaches scaling solution
- Correction to scaling is very large for realistic number of games
- Large variance may be due to small number of games

$$\sigma(t) = \frac{1/2 - q}{\sqrt{3}} + f(t)$$

Large!

Variance inadequate to characterize competitiveness!
The distribution of win percentage

- Treat $q$ as a fitting parameter, $\text{time} =$ number of games
- Allows to estimate $q_{\text{model}}$ for different leagues
The upset frequency

- Upset frequency as a measure of predictability
  \[ q = \frac{\text{Number of upsets}}{\text{Number of games}} \]

- Addresses the variability in the number of games
- Measure directly from game-by-game results
  - Ties: count as 1/2 of an upset (small effect)
  - Ignore games by teams with equal records
  - Ignore games by teams with no record
The upset frequency

<table>
<thead>
<tr>
<th>League</th>
<th>$q_1$</th>
<th>$q_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FA</td>
<td>0.452</td>
<td>0.459</td>
</tr>
<tr>
<td>MLB</td>
<td>0.441</td>
<td>0.413</td>
</tr>
<tr>
<td>NHL</td>
<td>0.414</td>
<td>0.383</td>
</tr>
<tr>
<td>NBA</td>
<td>0.365</td>
<td>0.316</td>
</tr>
<tr>
<td>NFL</td>
<td>0.364</td>
<td>0.309</td>
</tr>
<tr>
<td>ABL</td>
<td>???</td>
<td>0.392</td>
</tr>
</tbody>
</table>

$q$ differentiates the different sport leagues!

football, baseball most competitive
basketball, american football least competitive
Evolution with time

- Parity, predictability mirror each other
- **American football**, baseball increasing competitiveness
- Soccer decreasing competitiveness (past 60 years)

\[ \sigma = \frac{1/2 - q}{\sqrt{3}} \]

I. Discussion

• **Model limitation:** it does not incorporate
  - Game location: home field advantage
  - Game score
  - Upset frequency dependent on relative team strength
  - Unbalanced schedule

• **Model advantages:**
  - Simple, involves only 1 parameter
  - Enables quantitative analysis
1. Conclusions

- Parity characterized by variance in winning percentage
  - Parity measure requires standings data
  - Parity measure depends on season length

- Predictability characterized by upset frequency
  - Predictability measure requires game results data
  - Predictability measure independent of season length

- Two-team competition model allows quantitative modeling of sports competitions
2. Tournaments (post-season)
Single-elimination Tournaments

2006 NCAA Division I Men's Basketball Championship

First Round:

- Duke 1
- Southern U. 16
- G. Washington 8
- NC-Wilmington 9
- Syracuse 5
- Texas A&M 12
- LSU 4
- Iowa 11
- West Virginia 6
- Southern Ill. 11
- Iowa 2
- N. Western St. 11
- California 7
- N.C. State 10
- Texas 2
- Penn 15

Second Round:

- Hampton vs. Monmouth

Regional Sites:

- Atlanta
- Washington, D.C.
- Indianapolis
- Minneapolis

Regions:

- Midwest
- South
- East
- West

NCAA

2006 Final Four

Indianapolis

April 1

National Champion

*** ALL TIMES ARE LOCAL***

On March 12, the bracket committee will select two teams to play the opening round games March 14-16. The winning team will be a 16 seed in the region.

© 2006 National Collegiate Athletic Association. No commercial use without the NCAA's written permission.

The NCAA reserves all rights regarding the use of this bracket. The bracket should only be used for no-money, no-win contests, office pools or other gambling activities.

Binary Tree Structure
The competition model

- **Two teams play, loser is eliminated**
  \[ N \rightarrow N/2 \rightarrow N/4 \rightarrow \cdots \rightarrow 1 \]

- **Teams have inherent strength (or fitness) x**

- **Outcome of game depends on team strength**
  - Weaker team wins with probability \( q < 1/2 \)
  - Stronger team wins with probability \( 1 - q > 1/2 \)

\[
(x_1, x_2) \rightarrow \begin{cases} 
  x_1 & \text{probability } 1 - q \\
  x_2 & \text{probability } q 
\end{cases} \quad x_1 < x_2
\]
Recursive approach

- **Number of teams**
  \[ N = 2^k = 1, 2, 4, 8, \ldots \]

- **Cumulative probability distribution function for teams with fitness less than \( x \) to win an \( N \)-team tournament**

- **Closed equations for the cumulative distribution**
  \[ G_{2N}(x) = 2p G_N(x) + (1 - 2p) [G_N(x)]^2 \]

  **Nonlinear Recursion Equation**
Scaling properties

1. Scale of Winner

\[ x_* \sim N^{-\ln 2p/\ln 2} \]

2. Scaling Function

\[ G_N(x) \rightarrow \Psi \left( \frac{x}{x_*} \right) \]

3. Algebraic Tail

\[ 1 - \Psi(z) \sim z^{\ln 2p/\ln 2q} \]

1. Large tournaments produce strong winners
3. High probability for an upset
The scaling function

Universal shape

\[ \Psi(2pz) = 2p\Psi(z) + (1 - 2p)\Psi^2(z) \]

Broad tail

\[ \Psi'(z) \sim z^{\ln 2p/\ln 2q - 1} \]
College Basketball

- Teams ranked 1-16
- Well defined favorite
- Well defined underdog
- 4 winners each year
- Theory: $q=0.18$
- Simulation: $q=0.22$
- Data: $q=0.27$
- Data: 1978-2006
- 1600 games

2008: all four top seed advance; 1 in 150 chance!
Evolution, Men vs Women
2. Conclusions

- Strong teams fare better in large tournaments
- Tournaments can produce major upsets
- Distribution of winner relates parity with predictability
- Tournaments are efficient but not fair
3. Leagues
(regular season)
League champions

• N teams with fixed ranking
• In each game, favorite and underdog are well defined
• Favorite wins with probability $p > \frac{1}{2}$
  Underdog wins with probability $q < \frac{1}{2}$
• Each team plays $t$ games against random opponents
  - Regular random graph
• Team with most wins is the champion

How many games are needed for best team to win?
Random walk approach

- Probability team ranked $n$ wins a game
  $$P_n = p \frac{n-1}{N-1} + q \frac{N-n}{N-1}$$

- Number of wins performs a biased random walk
  $$w_n = P_n t \pm \sqrt{D_n t}$$

- Team $n$ can finish first at early times as long as
  $$(2p-1) \frac{n}{N} t \sim \sqrt{t}$$

- Rank of champion as function of $N$ and $t$
  $$n_* \sim \frac{N}{\sqrt{t}}$$
Length of season

• For best team to finish first
  \[ 1 \sim \frac{N}{\sqrt{t}} \]

• Each team must play
  \[ t \sim N^2 \]

• Total number of games
  \[ T \sim N^3 \]

1. Normal leagues are too short
2. Normal leagues: rank of winner \( \sim \sqrt{N} \)
3. League champions are a transient!
Distribution of outcomes

- Scaling distribution for the rank of champion
  \[ Q_n(t) \sim \frac{1}{n_*} \psi \left( \frac{n}{n_*} \right) \]
  \[ n_* \sim \frac{N}{\sqrt{t}} \]

- Probability worse team wins decays exponentially
  \[ Q_N(t) \sim \exp(-\text{const} \times t) \]

- Gaussian tail because
  \[ \psi \left( t^{1/2} \right) \sim \exp(-t) \]
  \[ \psi(z) \sim \exp (-\text{const} \times z^2) \]

- Normal league: Prob. (weakest team wins) \( \sim \exp(-N) \)

Leagues are fair: upset champions extremely unlikely
Leagues versus Tournaments

16 teams, $q=0.4$

$n_\ast \sim \sqrt{N}$

<table>
<thead>
<tr>
<th>$n$</th>
<th>league</th>
<th>tournament</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24.5</td>
<td>12.9</td>
</tr>
<tr>
<td>2</td>
<td>18.2</td>
<td>11.4</td>
</tr>
<tr>
<td>3</td>
<td>13.6</td>
<td>10.1</td>
</tr>
<tr>
<td>4</td>
<td>10.3</td>
<td>8.9</td>
</tr>
<tr>
<td>5</td>
<td>7.9</td>
<td>7.9</td>
</tr>
<tr>
<td>6</td>
<td>6.1</td>
<td>7.1</td>
</tr>
<tr>
<td>7</td>
<td>4.7</td>
<td>6.3</td>
</tr>
<tr>
<td>8</td>
<td>3.7</td>
<td>5.7</td>
</tr>
<tr>
<td>9</td>
<td>2.9</td>
<td>5.1</td>
</tr>
<tr>
<td>10</td>
<td>2.2</td>
<td>4.6</td>
</tr>
<tr>
<td>11</td>
<td>1.7</td>
<td>4.2</td>
</tr>
<tr>
<td>12</td>
<td>1.3</td>
<td>3.8</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>3.4</td>
</tr>
<tr>
<td>14</td>
<td>0.81</td>
<td>3.1</td>
</tr>
<tr>
<td>15</td>
<td>0.63</td>
<td>2.8</td>
</tr>
<tr>
<td>16</td>
<td>0.49</td>
<td>2.6</td>
</tr>
</tbody>
</table>
What is the likelihood the best team has best record?

<table>
<thead>
<tr>
<th>league</th>
<th>season</th>
<th>games</th>
<th>likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>NFL</td>
<td>short</td>
<td>predictable</td>
<td>30%</td>
</tr>
<tr>
<td>MLB*</td>
<td>long</td>
<td>random</td>
<td>31%</td>
</tr>
<tr>
<td>NHL</td>
<td>moderate</td>
<td>moderate</td>
<td>32%</td>
</tr>
<tr>
<td>NBA</td>
<td>moderate</td>
<td>predictable</td>
<td>45%</td>
</tr>
</tbody>
</table>

*90% likelihood requires 15000 games/team!!!
3. Conclusions

- Leagues are fair but inefficient
- Leagues do not produce major upsets
4. Ranking Algorithm
One preliminary round

- Preliminary round
  - Teams play a small number of games $T \sim N^t$
  - Top $M$ teams advance to championship round $M \sim N^\alpha$
  - Bottom $N-M$ teams eliminated
  - Best team must finish no worse than $M$ place $t \sim \frac{N^2}{M^2}$

- Championship round: plenty of games $T \sim M^3$

- Total number of games $T \sim N^{3-2\alpha} + N^{3\alpha}$

- Minimal when $M \sim N^{3/5}$, $T \sim N^{9/5}$
Two preliminary rounds

- Two stage elimination
  \[ N \rightarrow N^{\alpha_2} \rightarrow N^{\alpha_2\alpha_1} \rightarrow 1 \]

- Second round
  \[ T_2 \sim N^{3-2\alpha_2} + N^{\alpha_2(3-2\alpha_1)} + N^{3\alpha_1\alpha_2} \]

- Minimize number of games
  \[ 3 - 2\alpha_2 = \alpha_2(3 - 2\alpha_1) \quad \longrightarrow \quad \alpha_2 = \frac{15}{19} \]

- Further improvement in efficiency
  \[ T \sim N^{27/19} \]
Multiple preliminary rounds

- Each additional round further reduces $T$
  
  \[ T_k \sim N^{\gamma_k} \]
  \[ \gamma_k = \frac{1}{1 - (2/3)^{k+1}} \]

- Gradual elimination
  
  \[ N \rightarrow N^{\frac{57}{65}} \rightarrow N^{\frac{15}{19}} \rightarrow N^{\frac{3}{5}} \rightarrow 1 \]

- Teams play a small number of games initially

  Optimal linear scaling achieved using many rounds
  
  \[ T_\infty \sim N \]
  \[ M_\infty \sim N^{1/3} \]

  Preliminary elimination is very efficient!
4. Conclusions

• Gradual elimination is fair and efficient

• Preliminary rounds reduce the number of games

• In preliminary round, teams play a small number of games and almost all teams advance to next round
5. Social Dynamics
Competition and social dynamics

• Teams are agents
• Number of wins represents fitness or wealth
• Agents advance by competing against each other
• Competition is a mechanism for social differentiation
The social diversity model

- **Agents advance by competition**
  \[(i, j) \rightarrow \begin{cases} (i + 1, j) & \text{probability } p \\ (i, j + 1) & \text{probability } 1 - p \end{cases} \quad i > j\]

- **Agent decline due to inactivity**
  \[k \rightarrow k - 1 \quad \text{with rate } r\]

- **Rate equations**
  \[
  \frac{dG_k}{dt} = r(G_{k+1} - G_k) + pG_{k-1}(G_{k-1} - G_k) + (1 - p)(1 - G_k)(G_{k-1} - G_k) - \frac{1}{2}(G_k - G_{k-1})^2
  \]

- **Scaling equations**
  \[
  [(p + r - 1 + x) - (2p - 1)F(x)] \frac{dF}{dx} = 0
  \]
Organization into Social Structure

1. Middle class
   • Agents advance at different rates

2. Middle+lower class
   • Some agents advance at different rates
   • Some agents do not advance

3. Lower class
   • Agents do not advance

4. Egalitarian class
   • All agents advance at equal rates

Bonabeau 96
Concluding remarks

• Mathematical modeling of competitions sensible
• Minimalist models are a starting point
• Randomness a crucial ingredient
• Validation against data is necessary for predictive modeling
Publications

- Randomness in Competitions
  E. Ben-Naim, N.W. Hengartner

- Efficiency of Competitions
  E. Ben-Naim, N.W. Hengartner

- Scaling in Tournaments
  E. Ben-Naim, S. Redner, F. Vazquez
  Europhysics Letters 77, 30005 (2007)

- What is the Most Competitive Sport?
  E. Ben-Naim, F. Vazquez, S. Redner

- Dynamics of Multi-Player Games
  E. Ben-Naim, B. Kahng, and J.S. Kim

- On the Structure of Competitive Societies
  E. Ben-Naim, F. Vazquez, S. Redner

- Dynamics of Social Diversity
  E. Ben-Naim and S. Redner
Thanks

- Sidney Redner and Federico Vazquez
  Los Alamos & Boston University
- Jin Sup Kim and Byugnam Kahng
  Los Alamos & Seoul National University
- Nicholas Hengartner
  Los Alamos National Laboratory
- Micha Ben-Naim
  Massachusetts Institute of Technology