Parity and Predictability of Competitions: Nonlinear Dynamics of Sports

Eli Ben-Naim
Complex Systems Group & Center for Nonlinear Studies
Los Alamos National Laboratory

Federico Vazquez, Sidney Redner (Los Alamos & Boston University)

Talk, papers available from: http://cnls.lanl.gov/~ebn

2006 Dynamics Days Asia Pacific, Pohang, Korea, July 12, 2006
Plan

• Parity of sports leagues
• Theory: competition model
• Predictability of competitions
• Competition and social dynamics
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?
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
  \[ F(x) \]

In baseball
\[ 0.400 < x < 0.600 \]
\[ \sigma = 0.08 \]
# Data

- 300,000 Regular season games (all games)
- 5 Major sports leagues in US, 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>soccer</td>
<td>FA</td>
<td>Football Association</td>
<td>England</td>
<td>1888-2005</td>
<td>43,350</td>
</tr>
<tr>
<td>baseball</td>
<td>MLB</td>
<td>Major League Baseball</td>
<td>US</td>
<td>1901-2005</td>
<td>163,720</td>
</tr>
<tr>
<td>hockey</td>
<td>NHL</td>
<td>National Hockey League</td>
<td>US</td>
<td>1917-2005</td>
<td>39,563</td>
</tr>
<tr>
<td>basketball</td>
<td>NBA</td>
<td>National Basketball Association</td>
<td>US</td>
<td>1946-2005</td>
<td>43,254</td>
</tr>
<tr>
<td>american football</td>
<td>NFL</td>
<td>National Football League</td>
<td>US</td>
<td>1922-2004</td>
<td>11,770</td>
</tr>
</tbody>
</table>

Standard deviation in winning percentage

\[ F(x) \]

\( \text{data} \)
\( \text{theory} \)

NFL, NBA, NHL, MLB
Standard deviation in winning percentage

Distribution of winning percentage clearly distinguishes sports

- Baseball most competitive?
- American football least competitive?
Theory: competition model

- Two, randomly selected, teams play
- Outcome of game depends on team record
  - Better team wins with probability $1 - q$
  - Worst team wins with probability $q$
  - When two equal teams play, winner picked randomly
- Initially, all teams are equal (0 wins, 0 losses)
- Teams play once per unit time $\langle x \rangle = \frac{1}{2}$
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  worse team wins  equal 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  \( G_0 = 0 \)  \( G_\infty = 1 \)  Initial Conditions  \( G_k(t = 0) = 1 \)

Nonlinear Difference-Differential Equations
An exact solution

- **Winner 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**

\[
\frac{dP_k}{dt} = P_{k-1}
\]

- **Exact solution**

\[
G_k = \frac{1 + t + \frac{1}{2!} t^2 + \cdots + \frac{1}{k!} t^k}{1 + t + \frac{1}{2!} t^2 + \cdots + \frac{1}{(k+1)!} t^{k+1}}
\]
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**
  \[
  \frac{dG_k}{dt} = q(G_{k-1} - G_k) + (1/2 - q) (G_{k-1}^2 - G_k^2)
  \]
  
- **Treat number of wins as continuous**
  \[
  G_{k+1} - G_k \rightarrow \frac{\partial G}{\partial k}
  \]
  
- **Stationary distribution of winning percentage**
  \[
  G_k(t) \rightarrow F(x) \quad x = \frac{k}{t}
  \]

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

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

• Treat number of wins as continuous

Inviscid Burgers equation
\[
\frac{\partial v}{\partial t} + v \frac{\partial v}{\partial x} = 0
\]
\[
\frac{\partial G}{\partial t} + \left[ q + (1 - 2q)G \right] \frac{\partial G}{\partial k} = 0
\]

• 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

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

\[
\sigma(t) = \frac{1/2 - q}{\sqrt{3}} + f(t) \quad \text{Large!}
\]

Variance inadequate to characterize competitiveness!
The distribution of win percentage

- Treat $q$ as a fitting parameter, $\text{time}=\text{number of games}$
- 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
The upset frequency

<table>
<thead>
<tr>
<th>League</th>
<th>q</th>
<th>q_{model}</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>
</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
- Football decreasing competitiveness (past 60 years)

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

Football-American Football gap narrows from 9% to 2%!
All-time team records

- Provides the longest possible record ($t \sim 13000$)
- Close to a linear function

NY Yankees ($0.567$)

$\sigma = 0.024$

$q = 0.458$
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
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
Competition and Social Dynamics

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

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

- 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
  \[
  [(p + r - 1 + x) - (2p - 1)F(x)] \frac{dF}{dx} = 0
  \]
Social structures

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
• Parity and Predictability of Competitions  
  E. Ben-Naim, F. Vazquez, S. Redner  
• What is the Most Competitive Sport?  
  E. Ben-Naim, F. Vazquez, S. Redner  
  physics/0512143
• 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  
“I do not make predictions, especially not about the future.”

Yogi Bera