Sports as a Model for Competitive Societies

Eli Ben-Naim
Los Alamos National Laboratory

Perspectives in Nonlinear Science, UC San Diego, January 10, 2012

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

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
Los Alamos Middle School

Plan

1. Modeling competitions: how to use competition data
2. Tournaments: lose and you are out
3. Leagues: everybody competed with everybody
4. Ranking algorithm: how to rank fairly and efficiently
5. Modeling social dynamics

Motivation

- Evolution: species compete, fitter wins
- Society: people compete for social status
- Economics: companies compete for market share
- Arts, science, politics: awards, prizes, elections

Competition is everywhere

Why sports?

- Sports competition results are:
  - Accurate
  - Widely available
  - Complete

Sports as a laboratory for understanding competition

Theme

- Competitions are not perfectly predictable
- Outcome of a single competition is stochastic
- Winner of a series of competitions (league, tournament) is also subject to randomness

Randomness is inherent

1. Modeling competitions

What is the most competitive sport?

 Soccer

 Baseball

 Hockey

 Basketball

 Football

Can competitiveness be quantified?

What is the most competitive sport?

 Soccer

 Baseball

 Hockey

 Basketball

 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

$$F(x)$$

Major League Baseball
American League
2011 Season-end Standings

AMERICAN LEAGUE

East	W	L	PCT
y-New York Yankees	97	65	.599
w-Tampa Bay Rays	91	71	.562
Boston Red Sox	90	72	.556
Toronto Blue Jays	81	81	.500
Baltimore Orioles	69	93	.426

In baseball

$$0.400 < x < 0.600$$

$$\sigma = 0.08$$

Micha Ben-Naim
Los Alamos High School

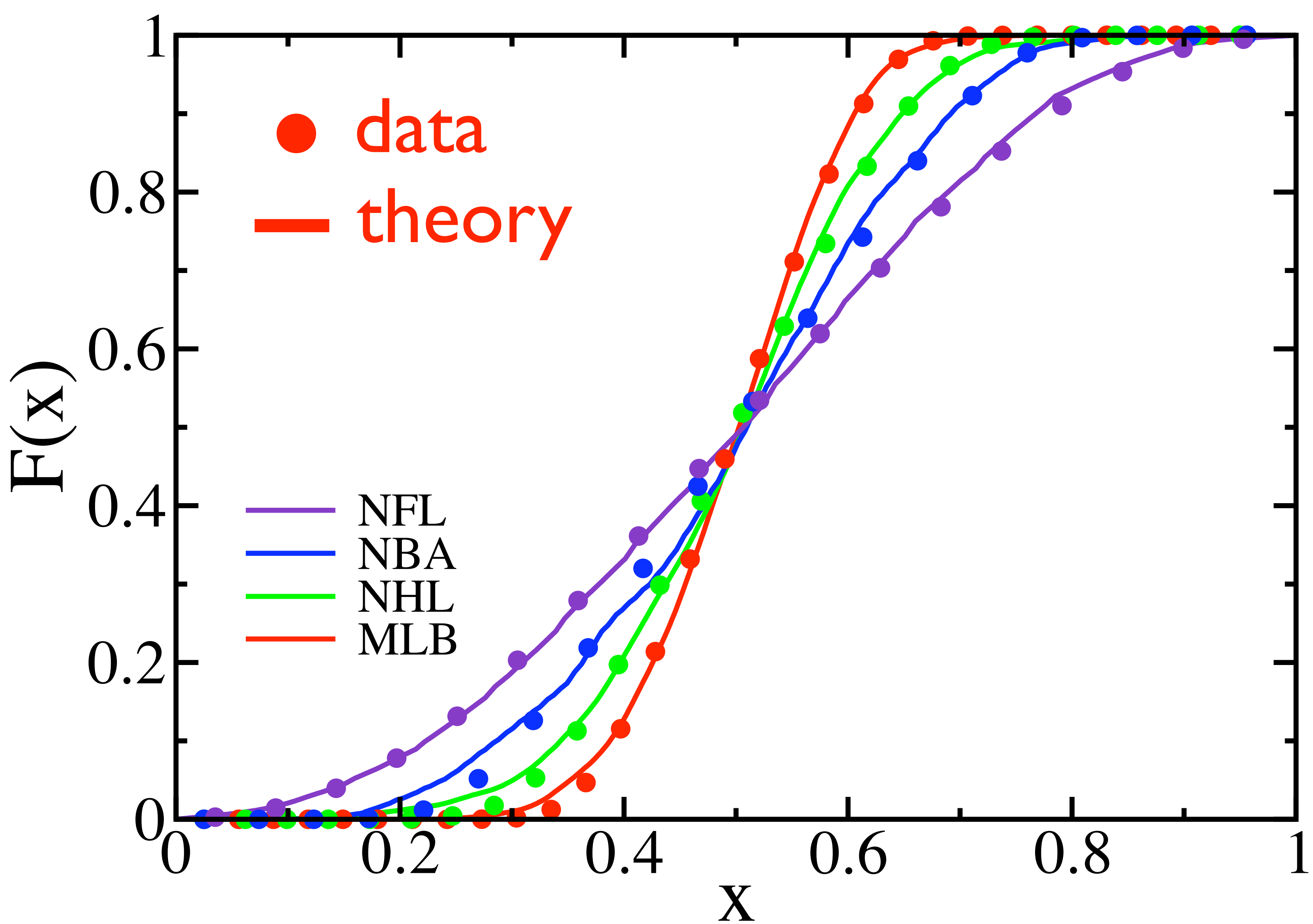
Data

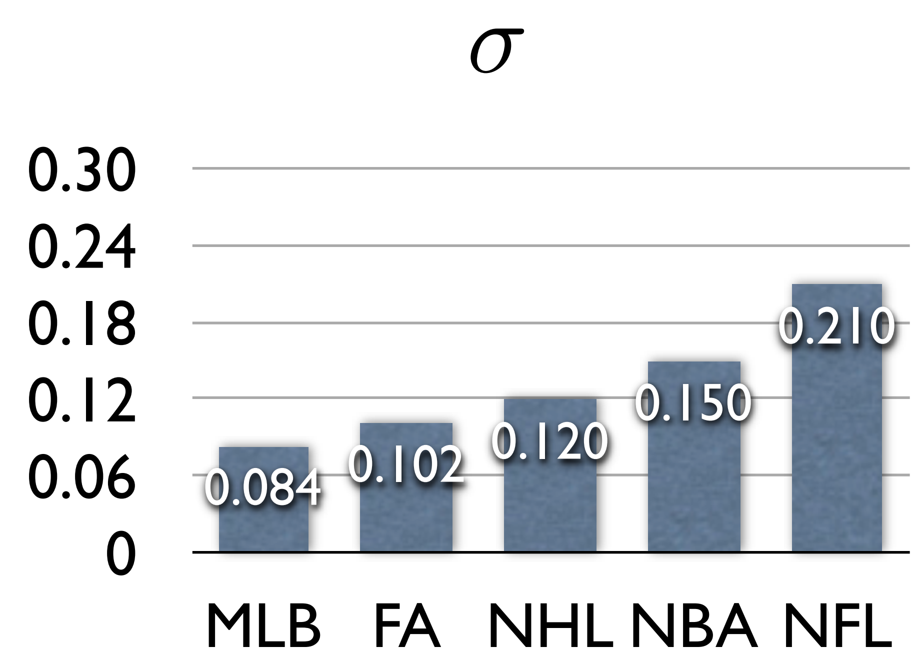
- 300,000 Regular season games (all games ever played)
- 5 Major sports leagues in United States & England

sport	league	full name	country	years	games
soccer	FA	Football Association		1888-2005	43,350
baseball	MLB	Major League Baseball		1901-2005	163,720
hockey	NHL	National Hockey League		1917-2005	39,563
basketball	NBA	National Basketball Association		1946-2005	43,254
football	NFL	National Football League		1922-2004	11,770

source: http://www.shrpsports.com/ http://www.the-english-football-archive.com/

Standard deviation in winning percentage





- Baseball most competitive?
- Football least competitive?

Distribution of winning percentage clearly distinguishes sports

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 $\longrightarrow \begin{cases} q = 1/2 & \text{random} \\ q = 0 & \text{deterministic} \end{cases}$
  - Stronger team wins with probability p>1/2 $\quad p + q = 1$

$$(i,j) \to \begin{cases} (i+1,j) & \text{probability } p \\ (i,j+1) & \text{probability } 1-p \end{cases} \qquad i > j$$

  - When two equal teams play, winner picked randomly
- Initially, all teams are equal (0 wins, 0 losses)
- Teams play once per unit time $\quad \langle x \rangle = \dfrac{1}{2}$

Rate equation approach

- Probability distribution functions

$$g_k = \text{fraction of teams with } k \text{ wins}$$

$$G_k = \sum_{j=0}^{k-1} g_j = \text{fraction of teams with less than } k \text{ wins} \qquad H_k = 1 - G_{k+1} = \sum_{j=k+1}^{\infty} g_j$$

- Evolution of the probability distribution

$$\frac{dg_k}{dt} = (1-q)(g_{k-1}G_{k-1} - g_k G_k) + q(g_{k-1}H_{k-1} - g_k H_k) + \frac{1}{2}\left(g_{k-1}^2 - g_k^2\right)$$

better team wins — worse team wins — equal teams play

- Closed equations for the cumulative distribution

$$\frac{dG_k}{dt} = q(G_{k-1} - G_k) + (1/2 - q)\left(G_{k-1}^2 - G_k^2\right)$$

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

Nonlinear Difference-Differential Equations

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

$$\frac{dP_k}{dt} = P_{k-1}$$

- 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

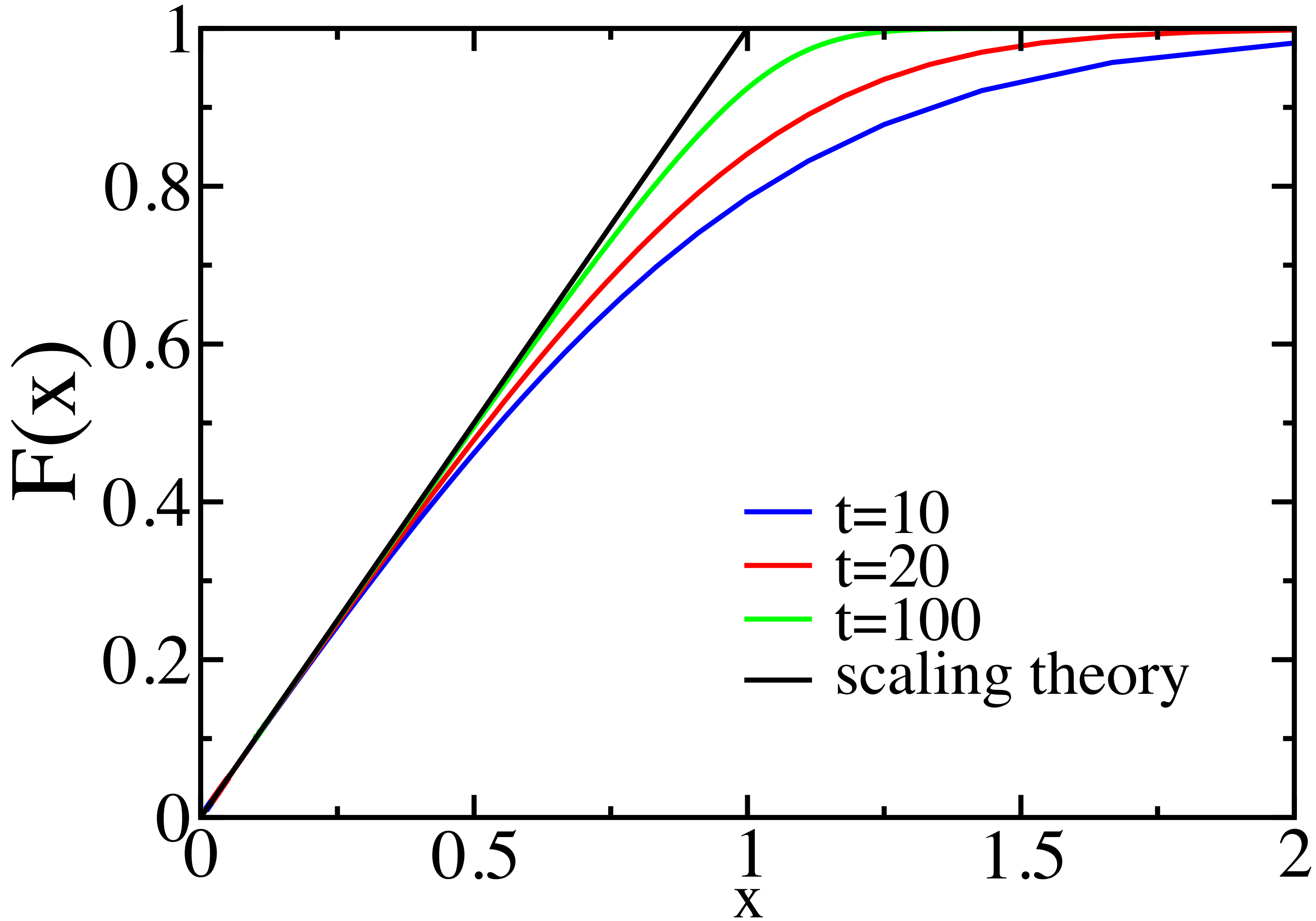
$$G_k \to \frac{k+1}{t}$$

- 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

$$\frac{dG_k}{dt} = q(G_{k-1} - G_k) + (1/2 - q)\left(G_{k-1}^2 - G_k^2\right)$$

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

Inviscid Burgers equation

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = 0$$

$$\frac{\partial G}{\partial t} + [q + (1-2q)G]\,\frac{\partial G}{\partial k} = 0$$

- Stationary distribution of winning percentage

$$G_k(t) \to F(x) \qquad x = \frac{k}{t}$$

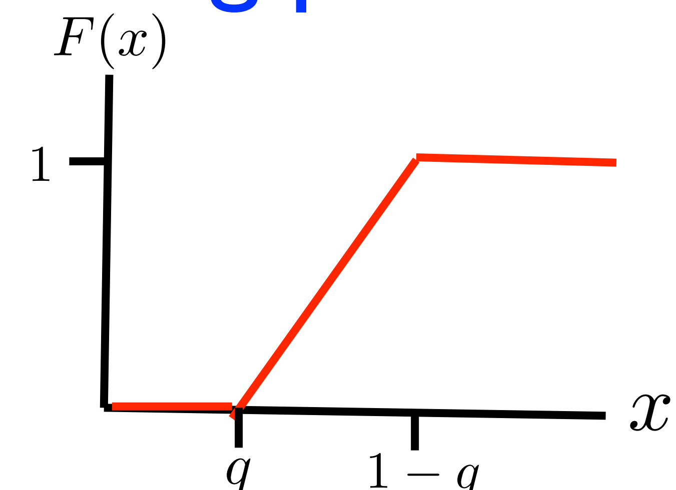
- Scaling equation

$$[(x-q) - (1-2q)F(x)]\,\frac{dF}{dx} = 0$$

Scaling solution

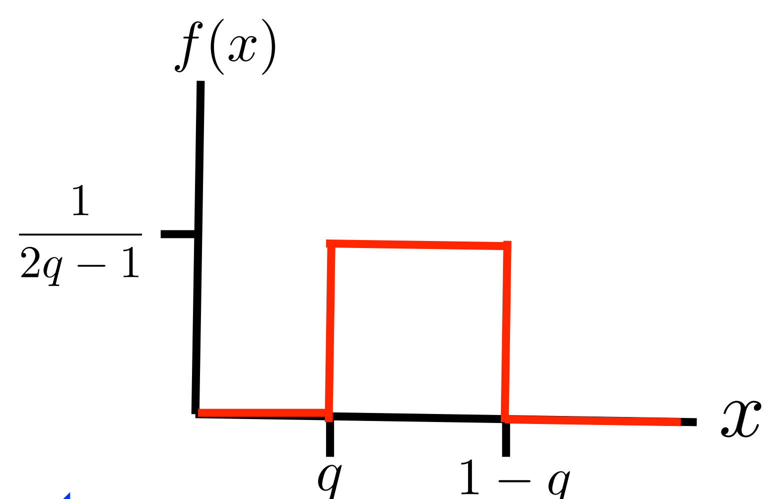
- Stationary distribution of winning percentage

$$F(x) = \begin{cases} 0 & 0 < x < q \\ \dfrac{x-q}{1-2q} & q < x < 1-q \\ 1 & 1-q < x. \end{cases}$$



- Distribution of winning percentage is uniform

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

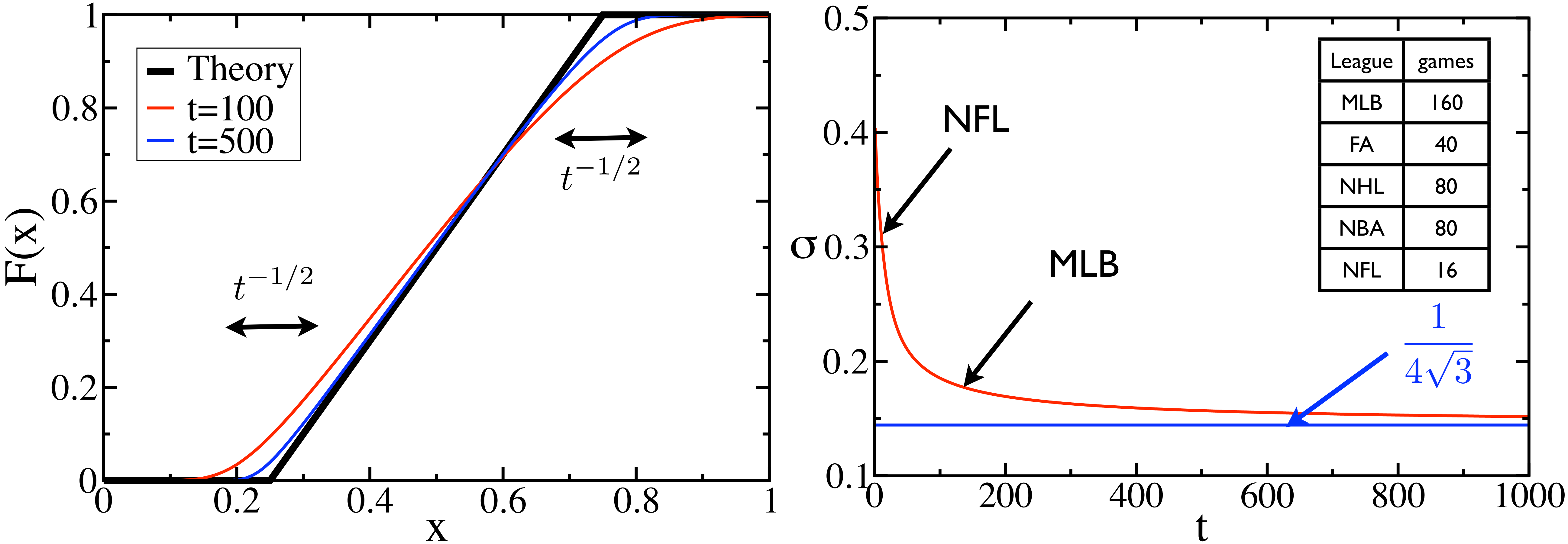


- Variance in winning percentage

$$\sigma = \frac{1/2 - q}{\sqrt{3}} \longrightarrow \begin{cases} q = 1/2 & \text{perfect parity} \\ q = 0 & \text{maximum disparity} \end{cases}$$

Approach to scaling

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



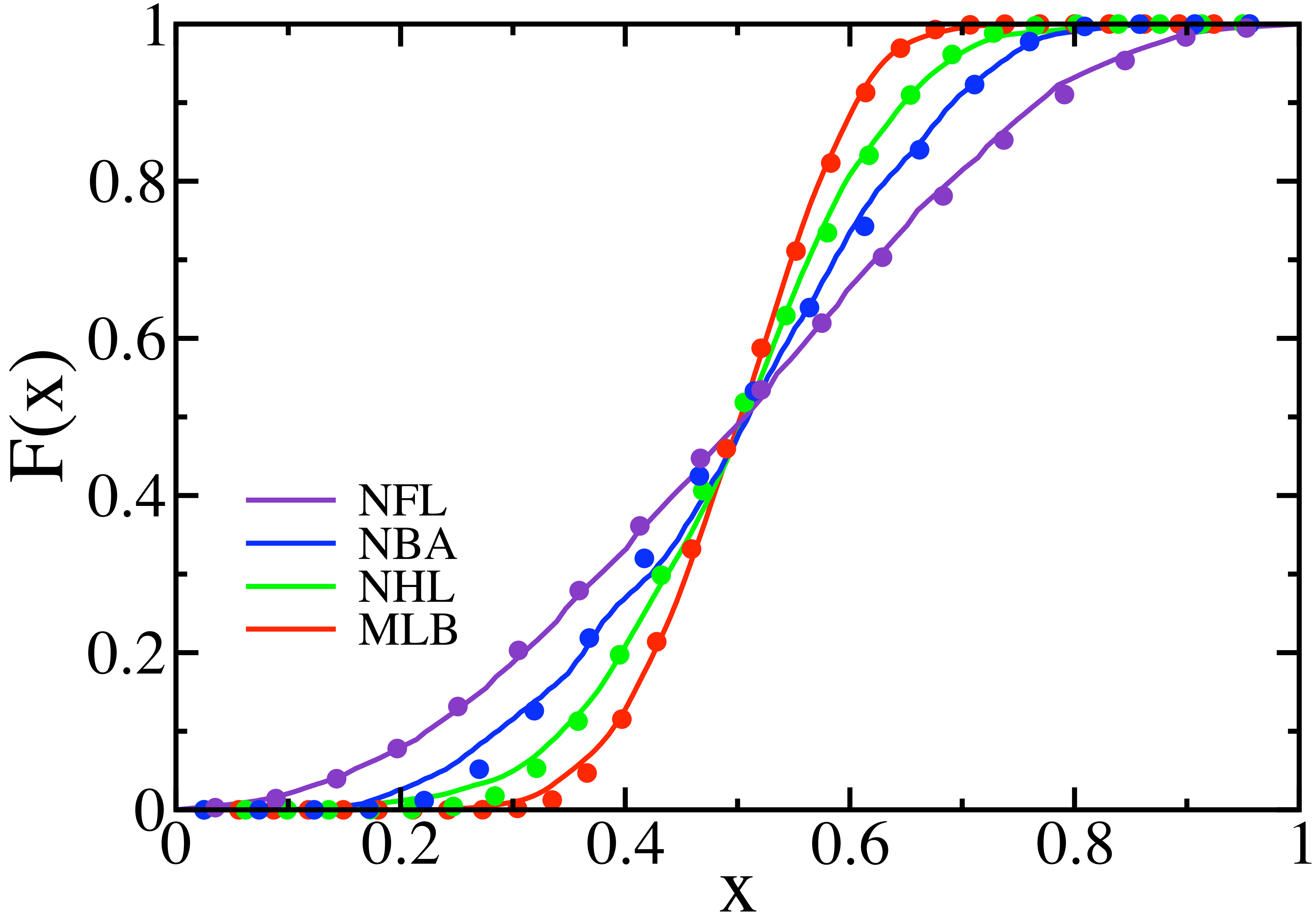
League	games
MLB	160
FA	40
NHL	80
NBA	80
NFL	16

- 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, time=number of games
- Allows to estimate q_{model} for different leagues

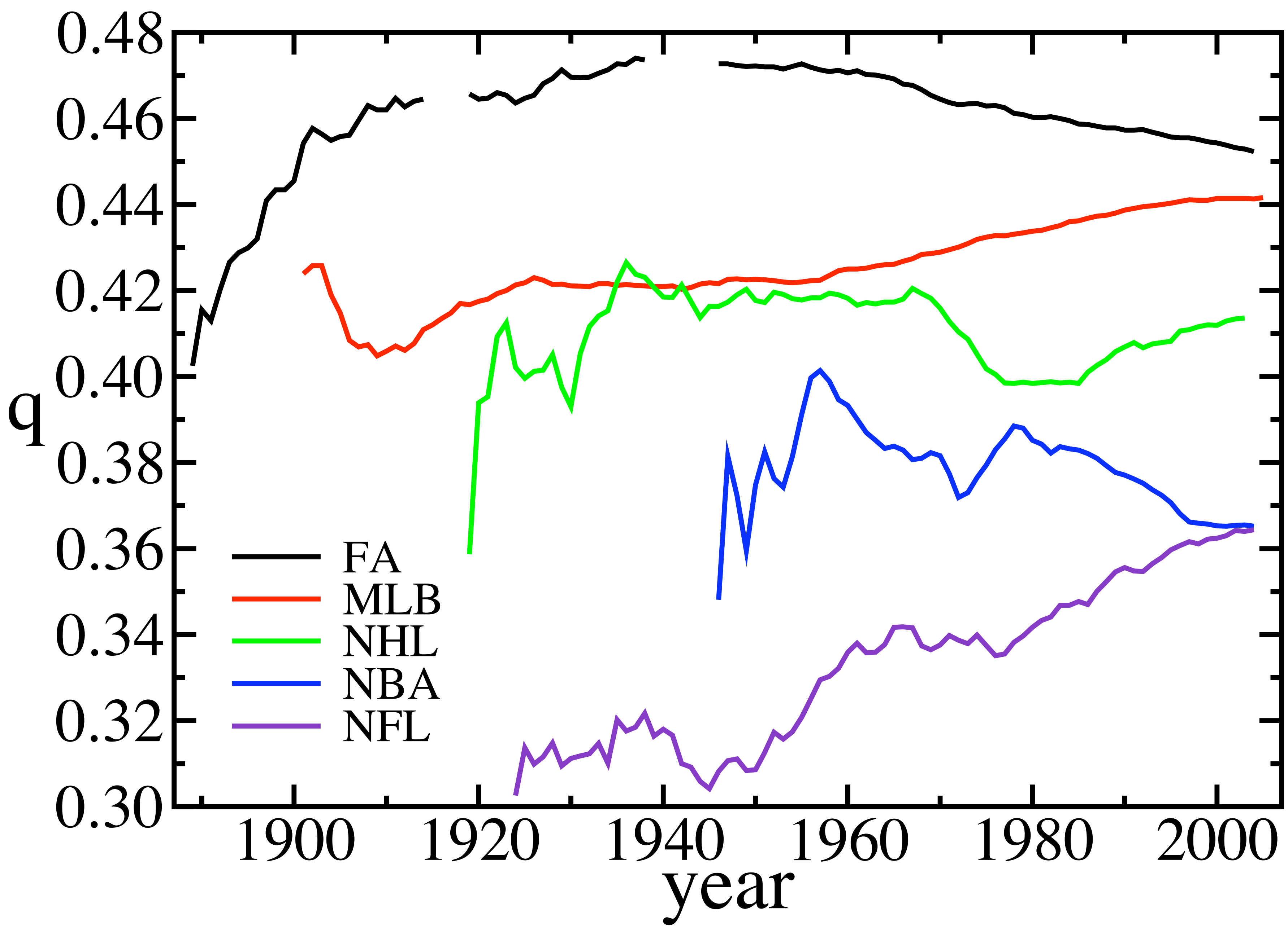
The upset frequency

- 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

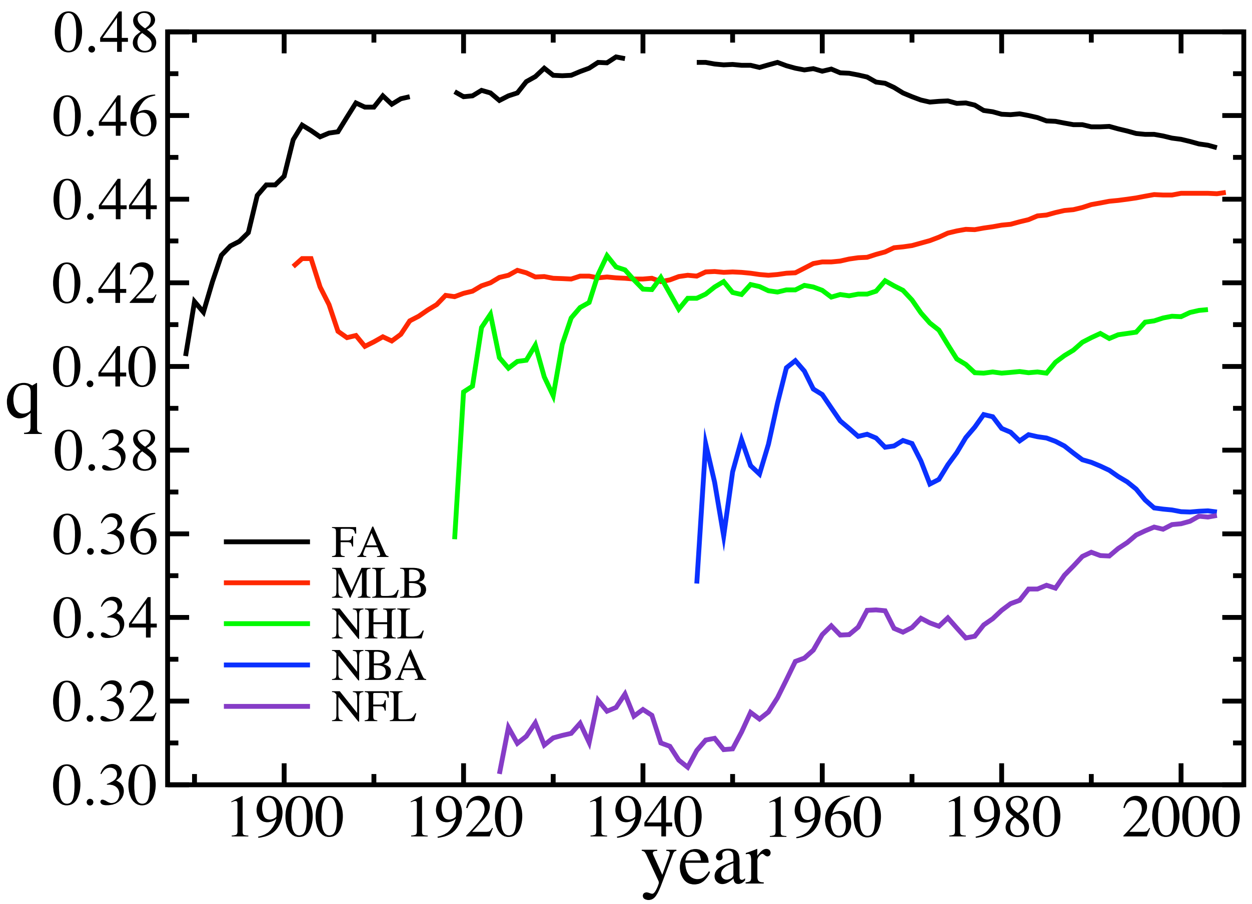


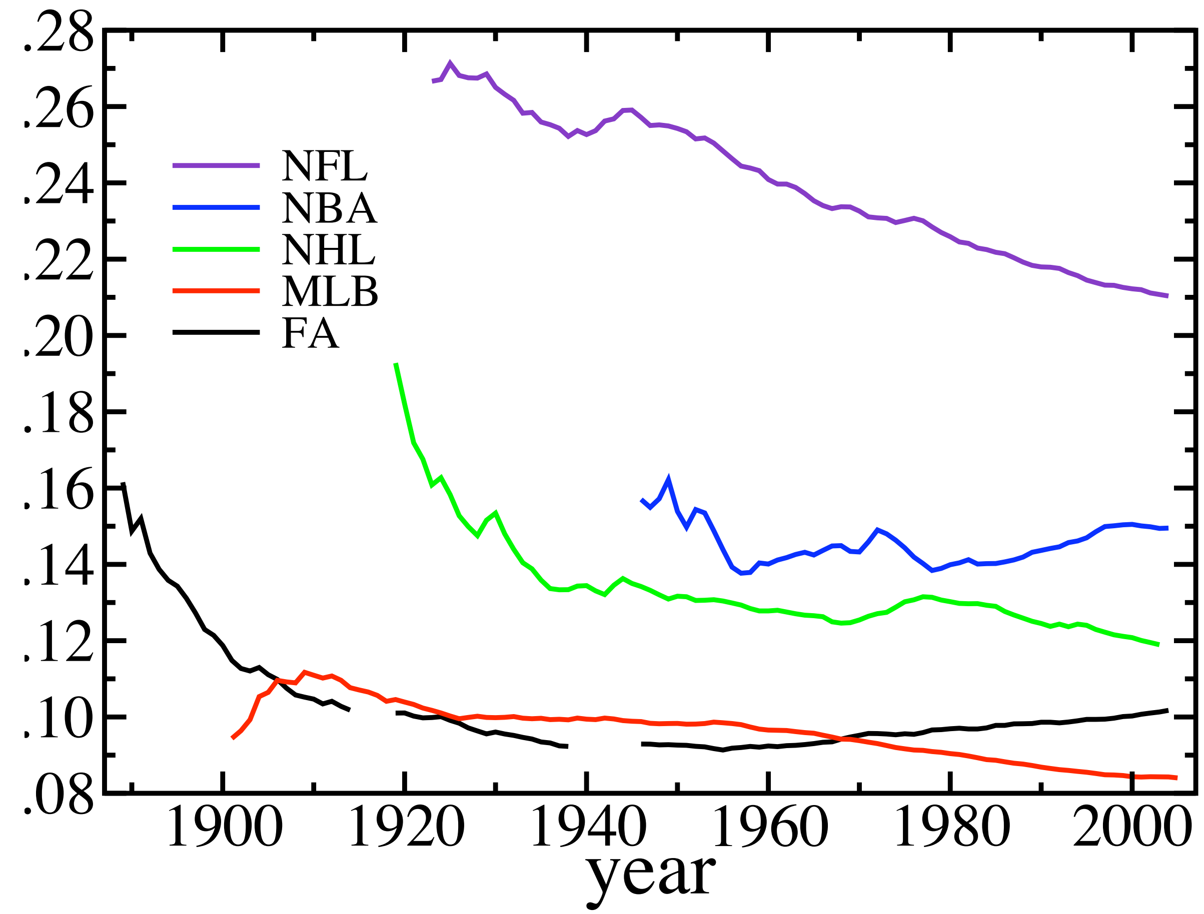
League	q	q_{model}
FA	**0.452**	0.459
MLB	**0.441**	0.413
NHL	**0.414**	0.383
NBA	**0.365**	0.316
NFL	**0.364**	0.309

q differentiates the different sport leagues!

Soccer, baseball most competitive
Basketball, football least competitive

Evolution with time





- Parity, predictability mirror each other $\sigma = \frac{1/2 - q}{\sqrt{3}}$
- Football, baseball increasing competitiveness
- Soccer decreasing competitiveness (past 60 years)

S.J. Gould, *Full House, The spread of excellence from Plato to Darwin*, 1996

I. Discussion

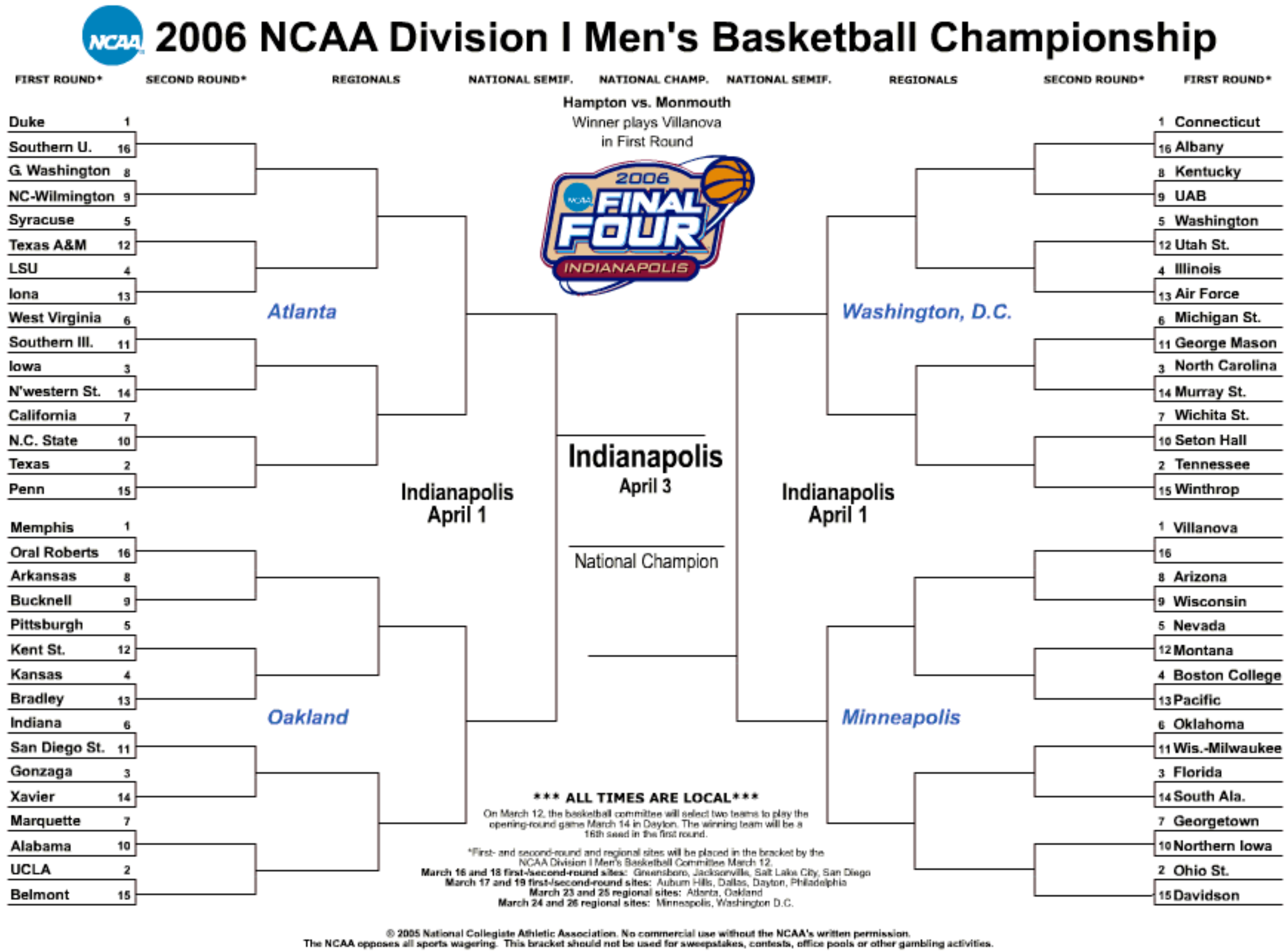
- Model limitation: it does not incorporate
  - Game location: home field advantage
  - Game score
  - Upset frequency dependent on relative team strength
  - Unbalanced schedule
- Model advantages:
  - Simple, involves only 1 parameter
  - Enables quantitative analysis

I. 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



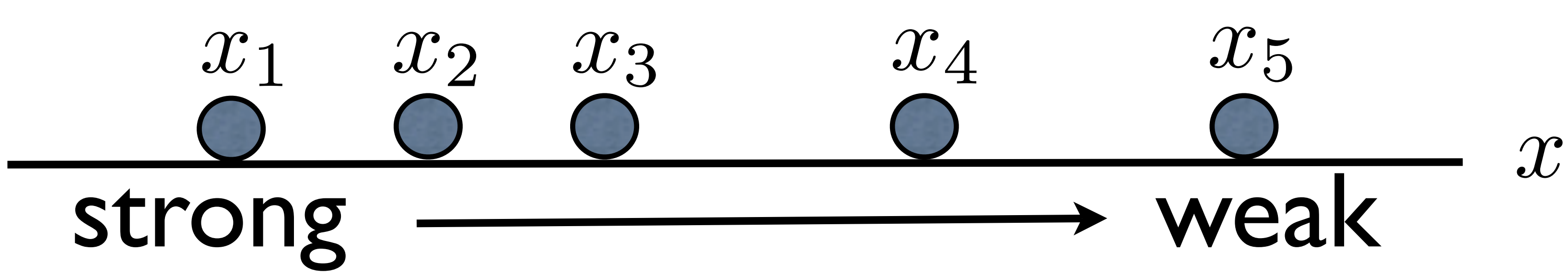
Binary Tree Structure

The competition model

- Two teams play, loser is eliminated

$$N \to N/2 \to N/4 \to \cdots \to 1$$

- Teams have inherent strength (or fitness) x



- Outcome of game depends on team strength

$$(x_1, x_2) \to \begin{cases} x_1 & \text{probability } 1-q \\ x_2 & \text{probability } q \end{cases} \qquad x_1 < x_2$$

Recursive approach

- Number of teams

$$N = 2^k = 1, 2, 4, 8, \ldots$$

- $G_N(x)$ = 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

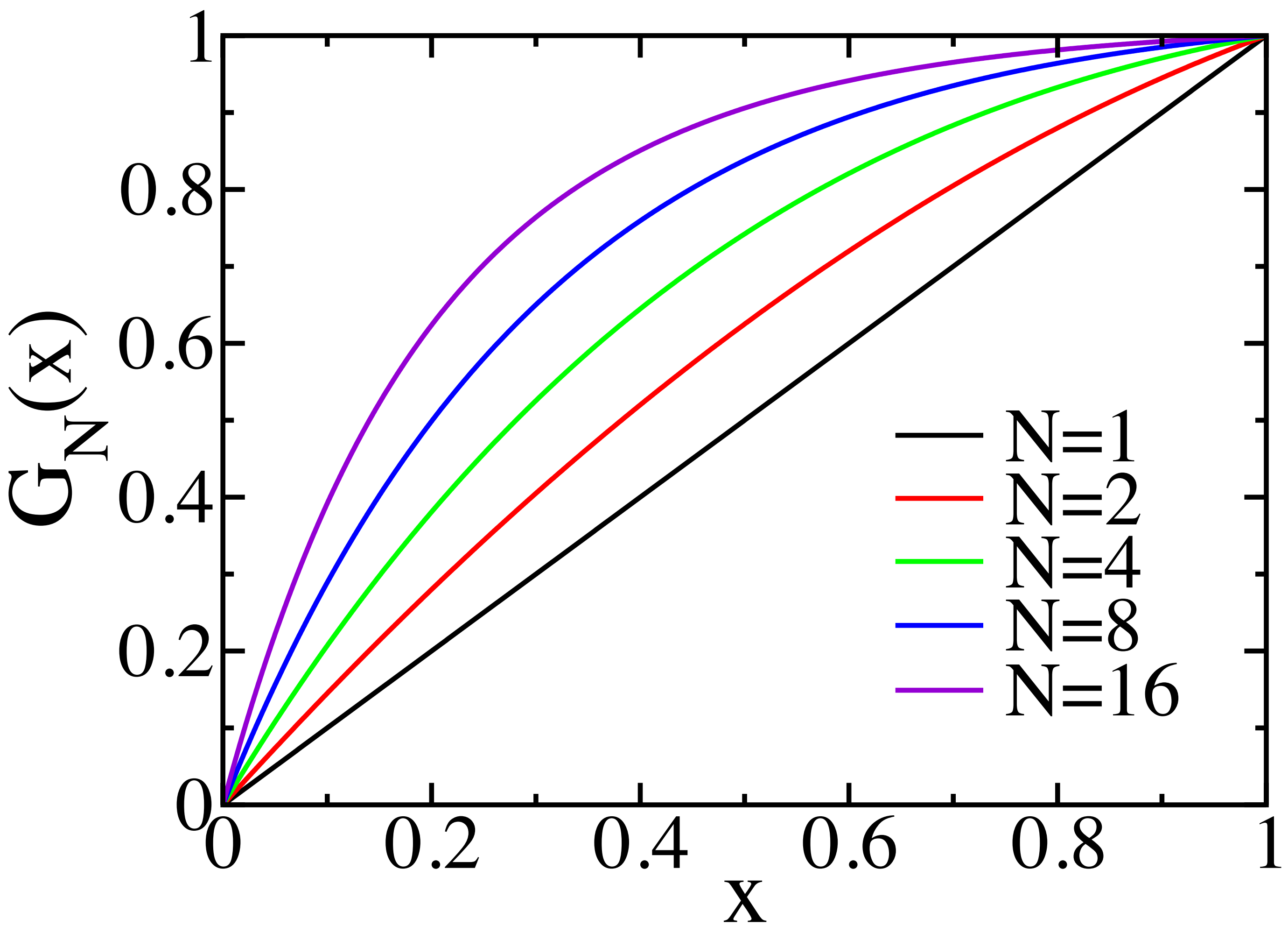
$$x_* \sim N^{-\ln 2p/\ln 2}$$

2. Scaling Function

$$G_N(x) \to \Psi\left(x/x_*\right)$$

3. Algebraic Tail

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

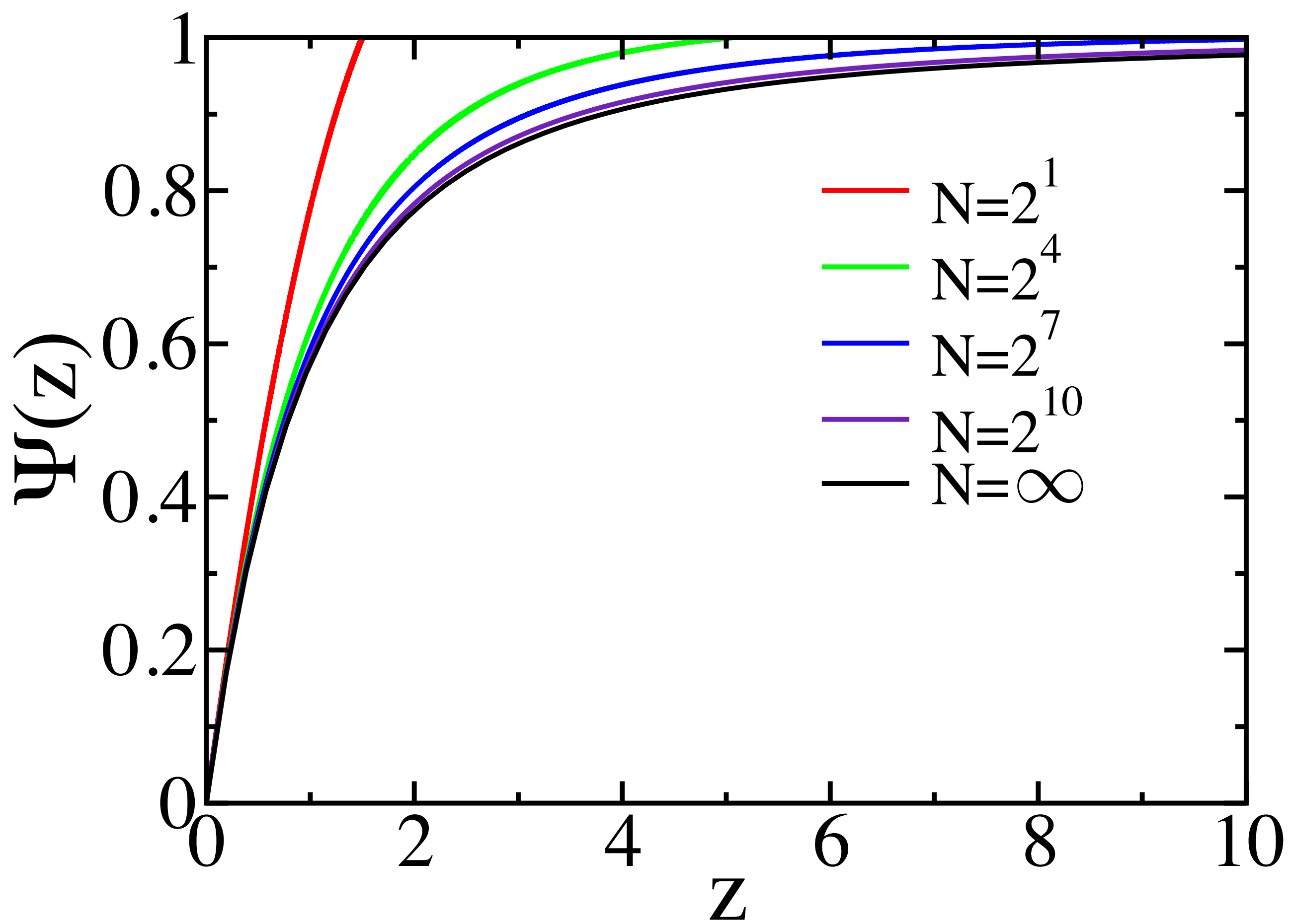


1. Large tournaments produce strong winners
3. High probability for an upset

The scaling function

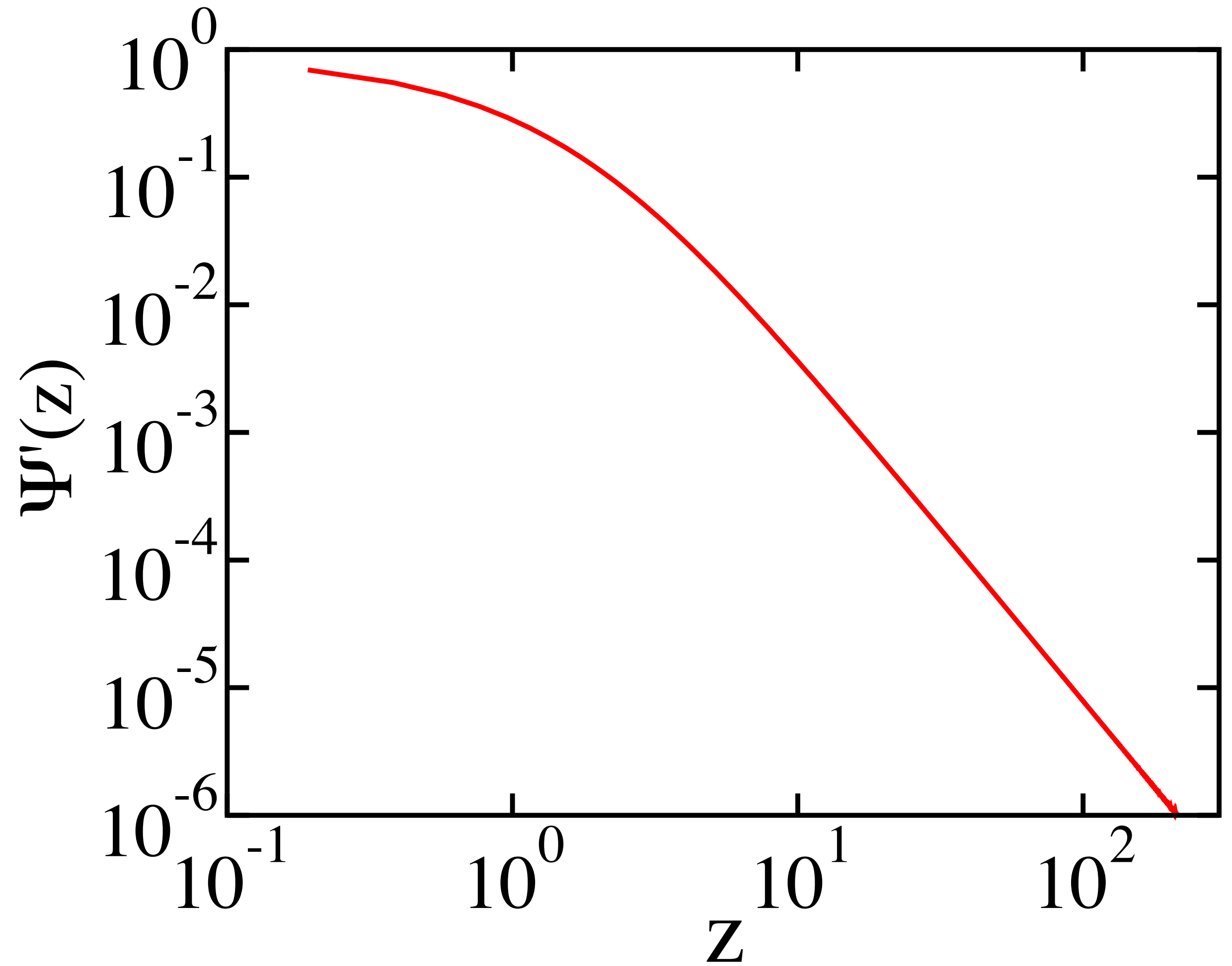
Universal shape

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

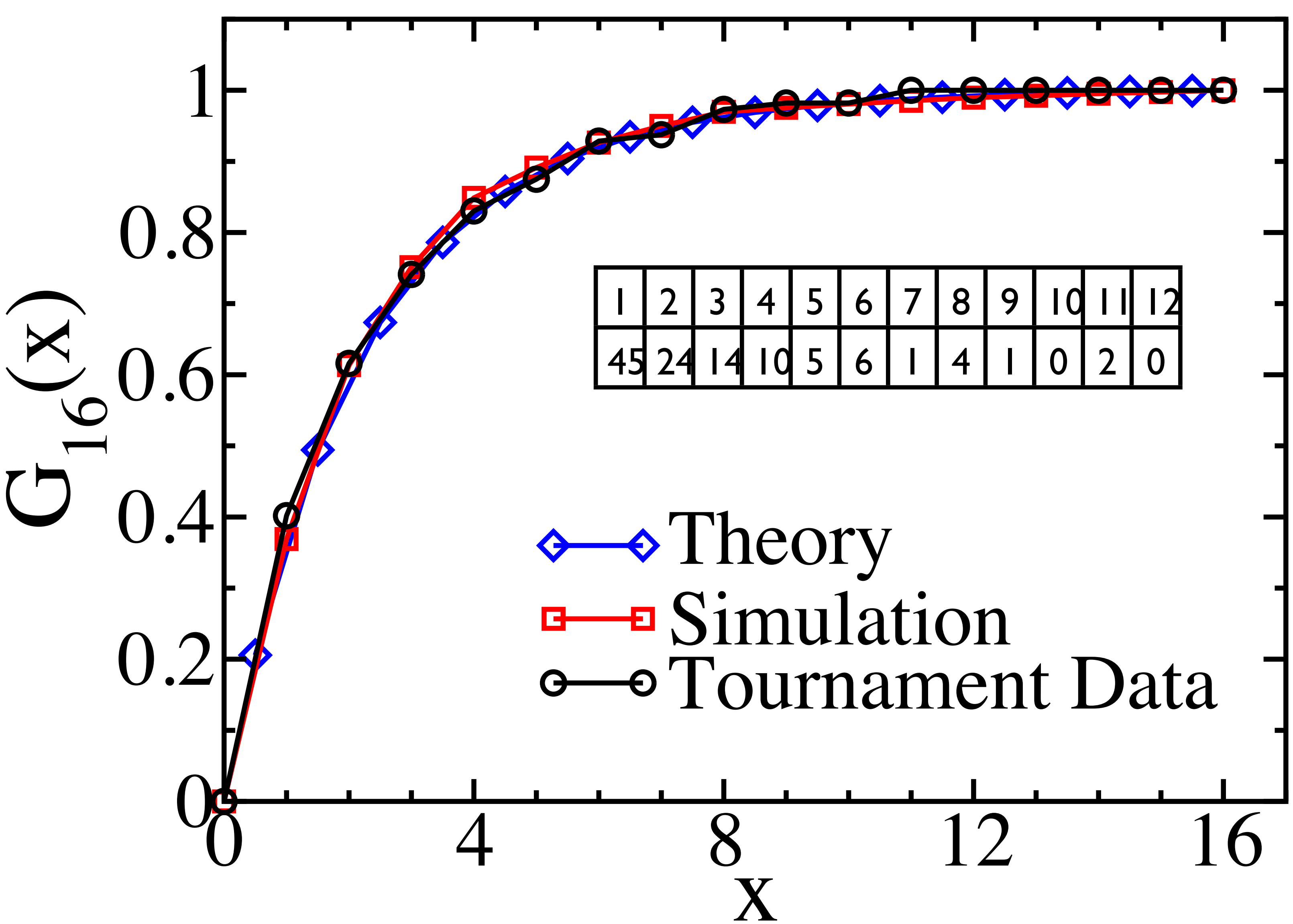


Broad tail

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



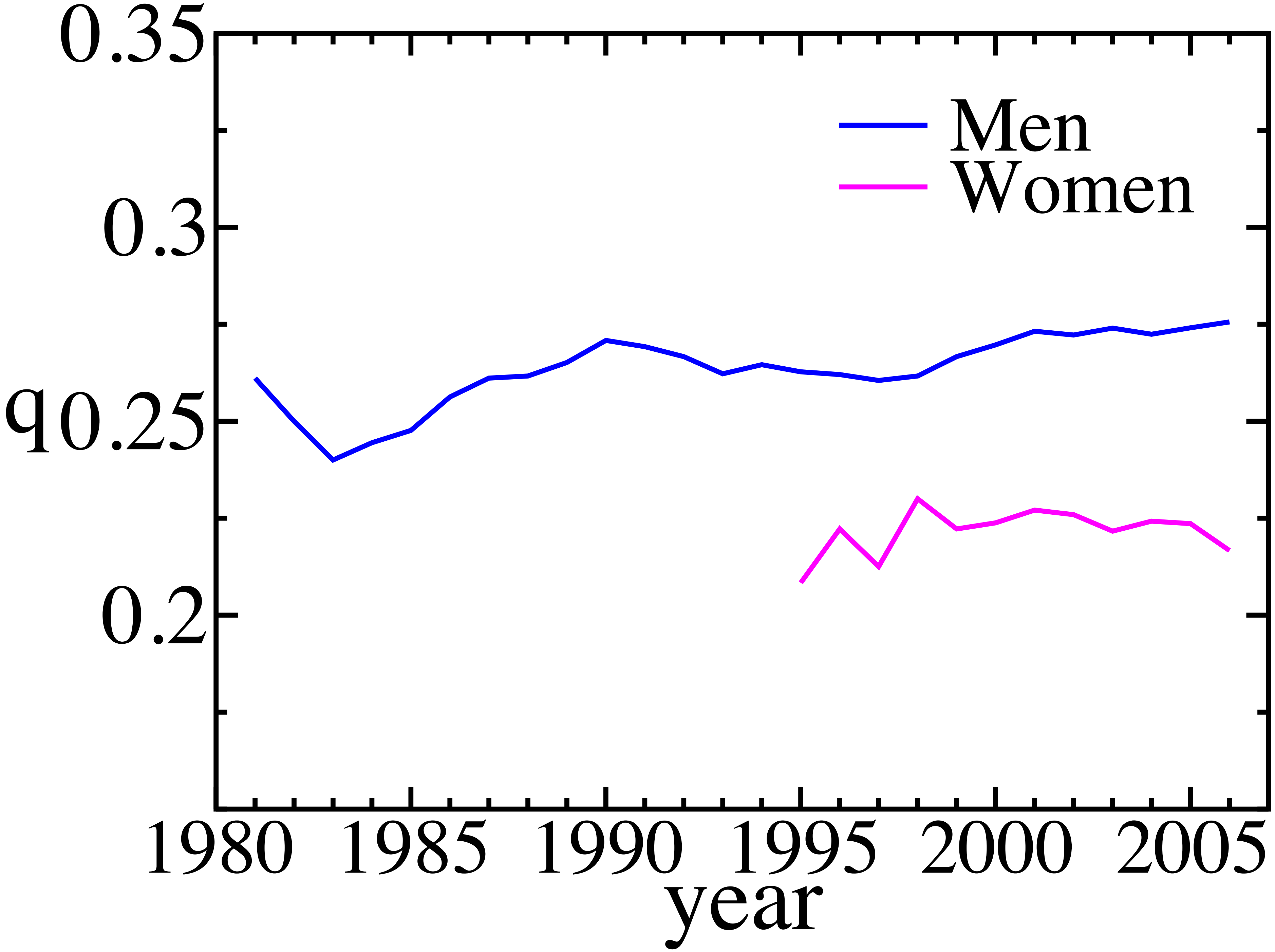
College Basketball



- 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



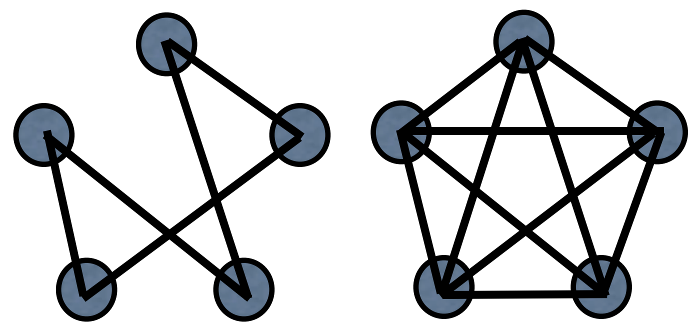
2. Conclusions

- 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

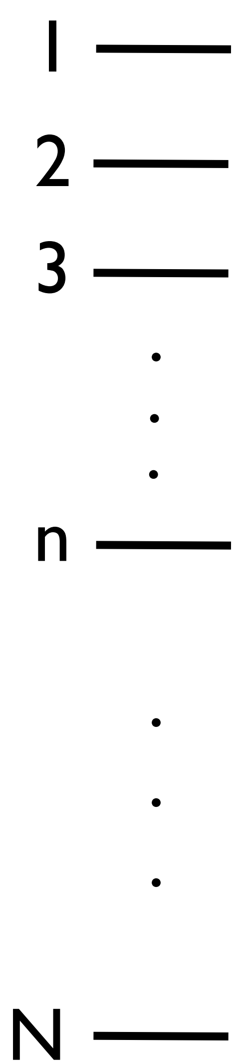
- N teams with fixed ranking
- In each game, favorite and underdog are well defined
- Favorite wins with probability p>1/2
  Underdog wins with probability q<1/2 $p + q = 1$
- Each team plays t games against random opponents
  - Regular random graph



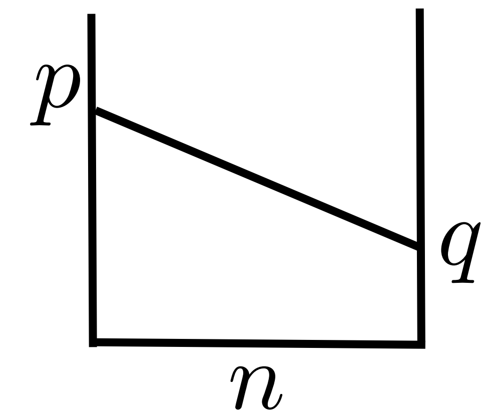
- Team with most wins is the champion

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

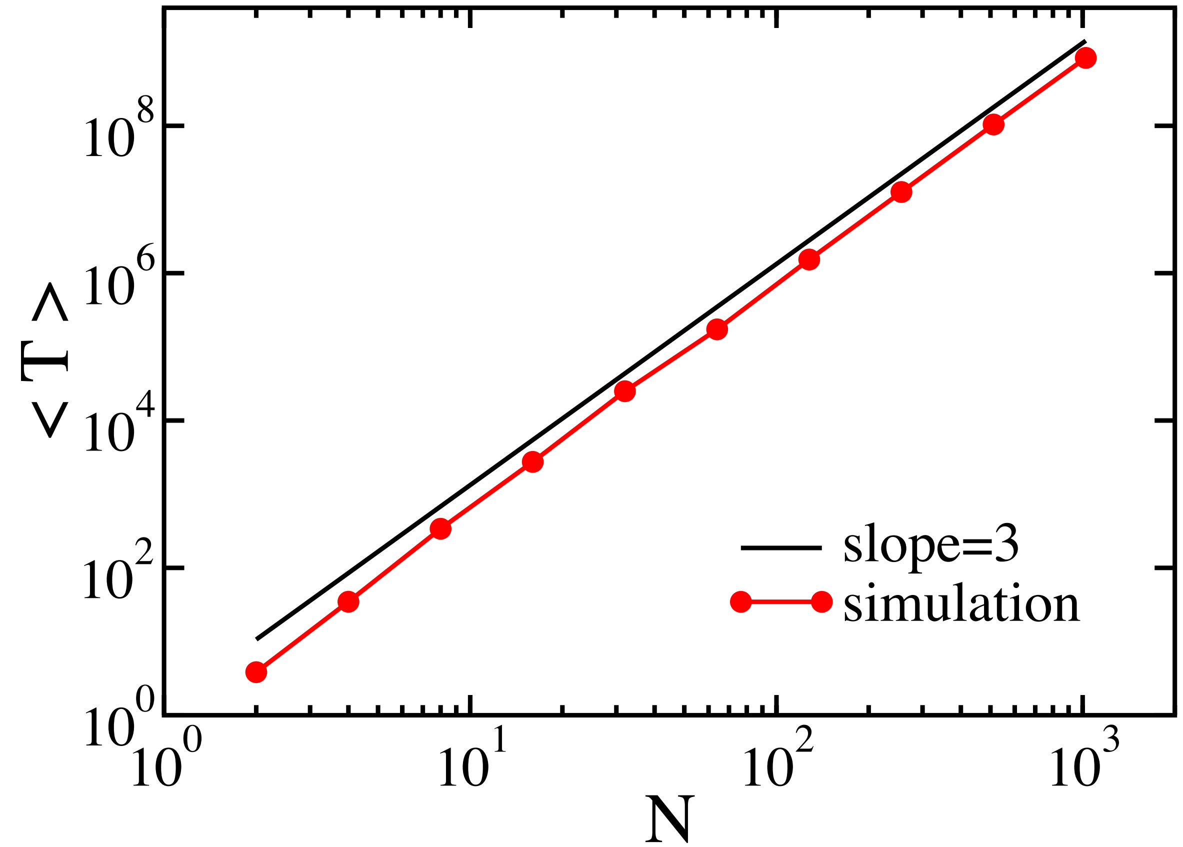
$$1 \sim \frac{N}{\sqrt{t}}$$

- 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

$$Q_n(t) \sim \frac{1}{n_*}\psi\left(\frac{n}{n_*}\right) \qquad n_* \sim \frac{N}{\sqrt{t}}$$

- Probability worse team wins decays exponentially

$$Q_N(t) \sim \exp(-\text{const} \times t)$$

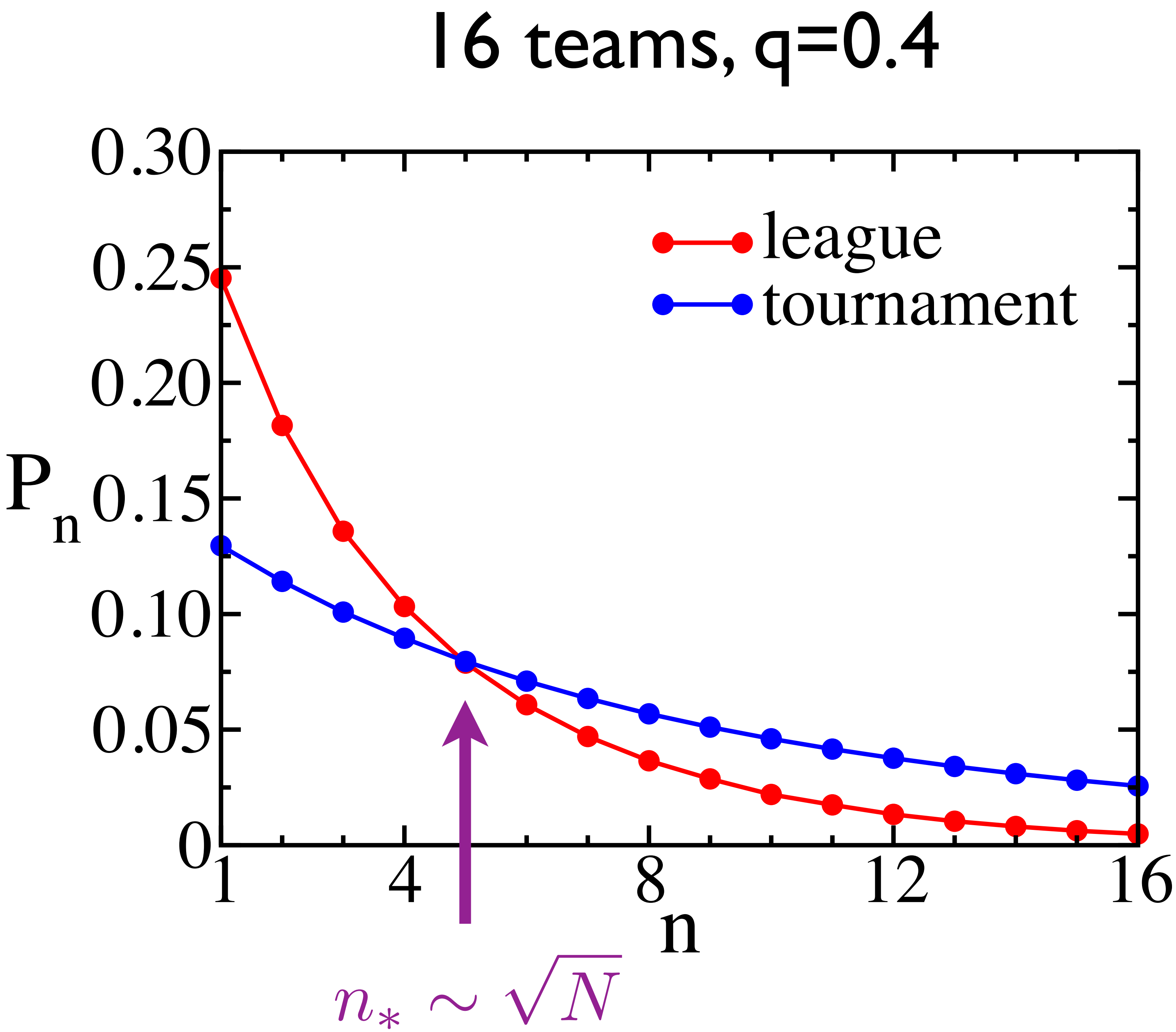
- Gaussian tail because $\psi\left(t^{1/2}\right) \sim \exp(-t)$

$$\psi(z) \sim \exp\left(-\text{const} \times z^2\right)$$

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

Leagues are fair: upset champions extremely unlikely

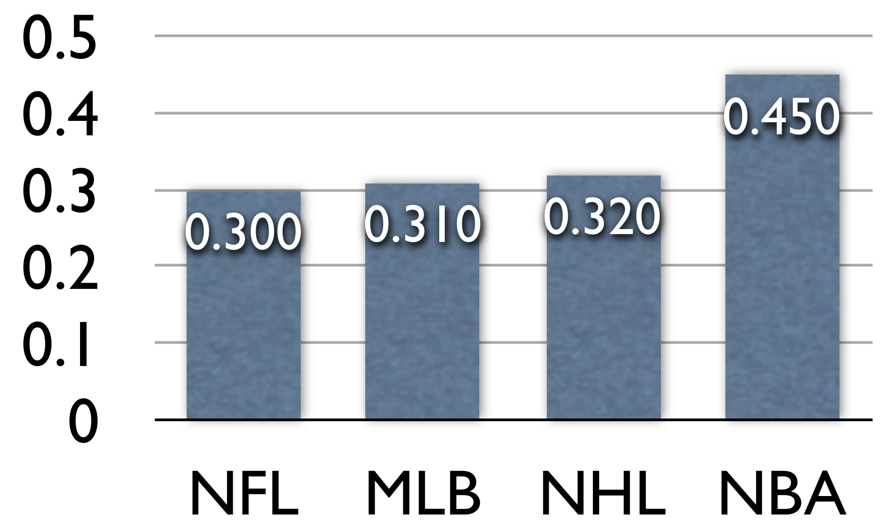
Leagues versus Tournaments



n	league	tournament
1	24.5	12.9
2	18.2	11.4
3	13.6	10.1
4	10.3	8.9
5	7.9	7.9
6	6.1	7.1
7	4.7	6.3
8	3.7	5.7
9	2.9	5.1
10	2.2	4.6
11	1.7	4.2
12	1.3	3.8
13	1.0	3.4
14	0.81	3.1
15	0.63	2.8
16	0.49	2.6

What is the likelihood the best team has best record?

league	season	games	likelihood
NFL	short	predictable	30%
MLB*	long	random	31%
NHL	moderate	moderate	32%
NBA	moderate	predictable	45%



*90% likelihood requires 15000 games/team!!!

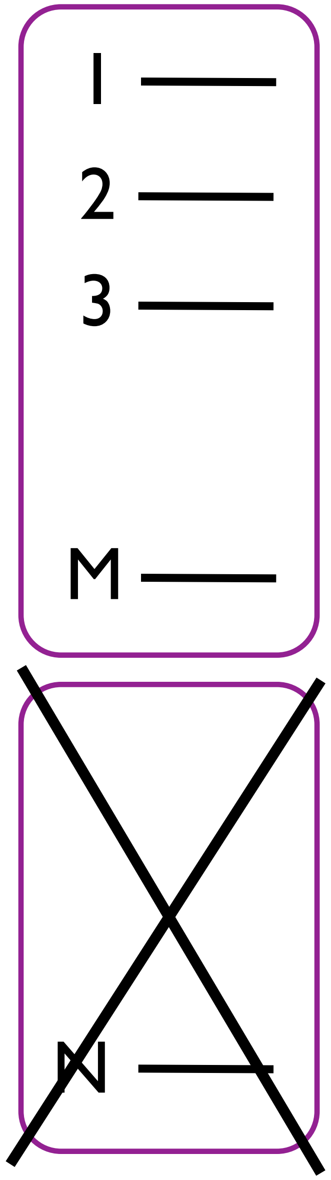
Interplay between length of season and predictability of games

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 $t \sim \dfrac{N^2}{M^2}$
- Championship round: plenty of games $T \sim M^3$
- Total number of games

$$T \sim N^{3-2\alpha} + N^{3\alpha}$$

- Minimal when

$$M \sim N^{3/5} \qquad T \sim N^{9/5}$$

Two preliminary rounds

- Two stage elimination

$$N \to N^{\alpha_2} \to N^{\alpha_2 \alpha_1} \to 1$$

- 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

$$3 - 2\alpha_2 = \alpha_2(3 - 2\alpha_1) \qquad \longrightarrow \qquad \alpha_2 = \frac{15}{19}$$

- Further improvement in efficiency

$$T \sim N^{27/19}$$

Multiple preliminary rounds

- Each additional round further reduces T

$$T_k \sim N^{\gamma_k} \qquad \gamma_k = \frac{1}{1-(2/3)^{k+1}}$$

- Gradual elimination

$$\gamma_k = 3, \frac{9}{5}, \frac{27}{19}, \frac{81}{65}, \cdots$$

$$N \to N^{\frac{57}{65}} \to N^{\frac{57}{65}\frac{15}{19}} \to N^{\frac{57}{65}\frac{15}{19}\frac{3}{5}} \to 1$$

- Teams play a small number of games initially

Optimal linear scaling achieved using many rounds

$$T_\infty \sim N \qquad M_\infty \sim N^{1/3}$$

optimal size of playoffs!

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

$$(i,j) \to \begin{cases} (i+1,j) & \text{probability } p \\ (i,j+1) & \text{probability } 1-p \end{cases} \qquad i > j$$

- Agent decline due to inactivity

$$k \to k-1 \qquad \text{with rate } r$$

- 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

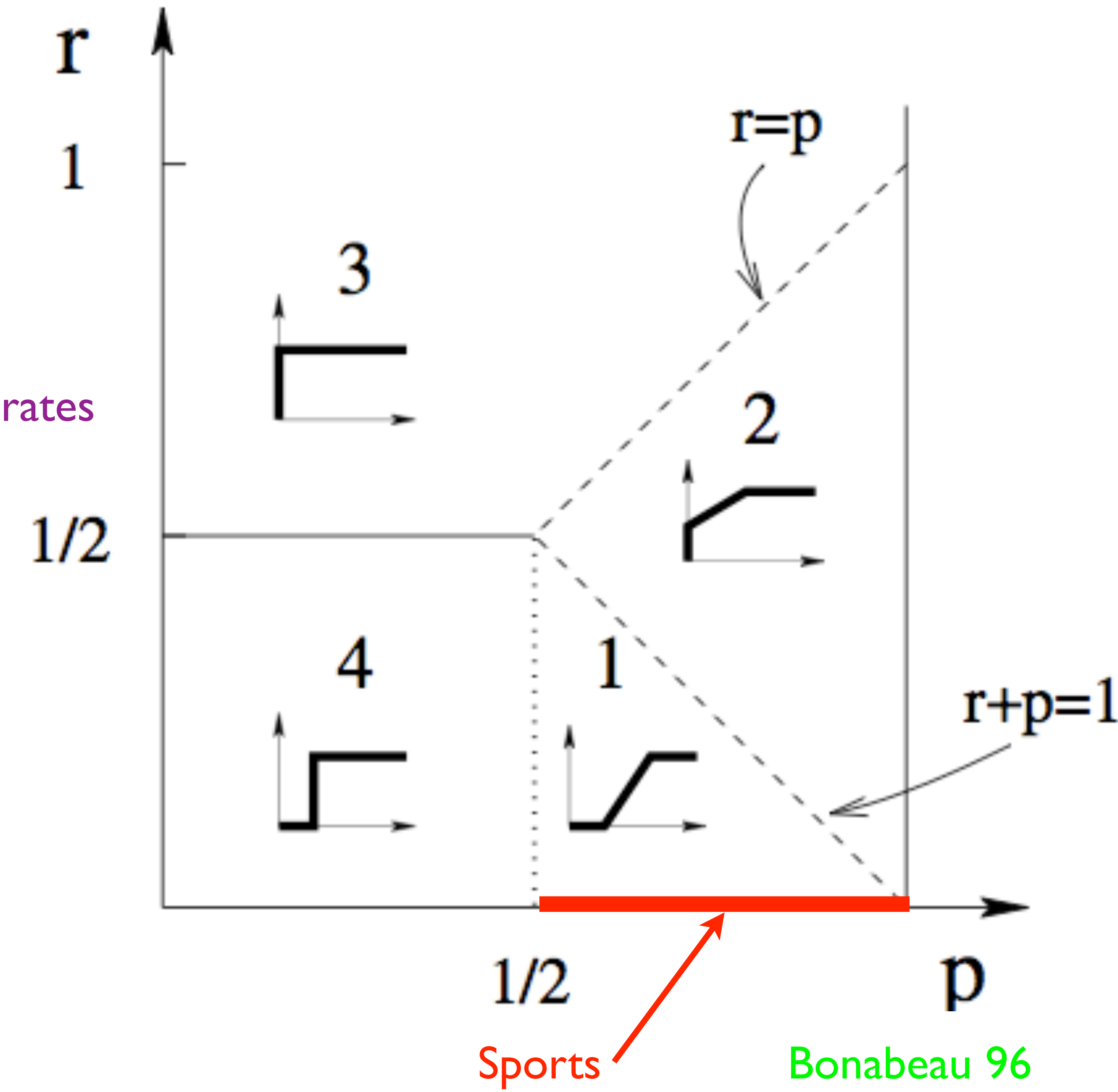
    Some agents do not advance

3. Lower class

    Agents do not advance

4. Egalitarian class

    All agents advance at equal rates



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

- Efficiency of Competitions
  E. Ben-Naim, N.W. Hengartner
  Phys. Rev. E **76**, 026106 (2007)
- 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
  J. Korean Phys. Soc. **50**, 124 (2007)
- Dynamics of Multi-Player Games
  E. Ben-Naim, B. Kahng, and J.S. Kim
  J. Stat. Mech. P07001 (2006)
- On the Structure of Competitive Societies
  E. Ben-Naim, F. Vazquez, S. Redner
  Eur. Phys. Jour. B **26** 531 (2006)
- Dynamics of Social Diversity
  E. Ben-Naim and S. Redner
  J. Stat. Mech. L11002 (2005)

"Prediction is very difficult,
especially about the future."

Niels Bohr

"Everything should be made as simple as possible but not simpler"

Freeman Dyson