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

Los Alamos
NATIONAL LABORATORY
EST.1943

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

$$F(x)$$

Major League Baseball
American League
2005 Season-end Standings

East	W	L	PCT
Boston	95	67	.586
New York	95	67	.586
Toronto	80	82	.494
Baltimore	74	88	.457
Tampa Bay	67	95	.414
Central	**W**	**L**	**PCT**
Chicago	99	63	.611
Cleveland	93	69	.574
Minnesota	83	79	.512
Detroit	71	91	.438
Kansas City	56	106	.346
West	**W**	**L**	**PCT**
Los Angeles	95	67	.586
Oakland	88	74	.543
Texas	79	83	.488
Seattle	69	93	.426

In baseball

$$0.400 < x < 0.600$$

$$\sigma = 0.08$$

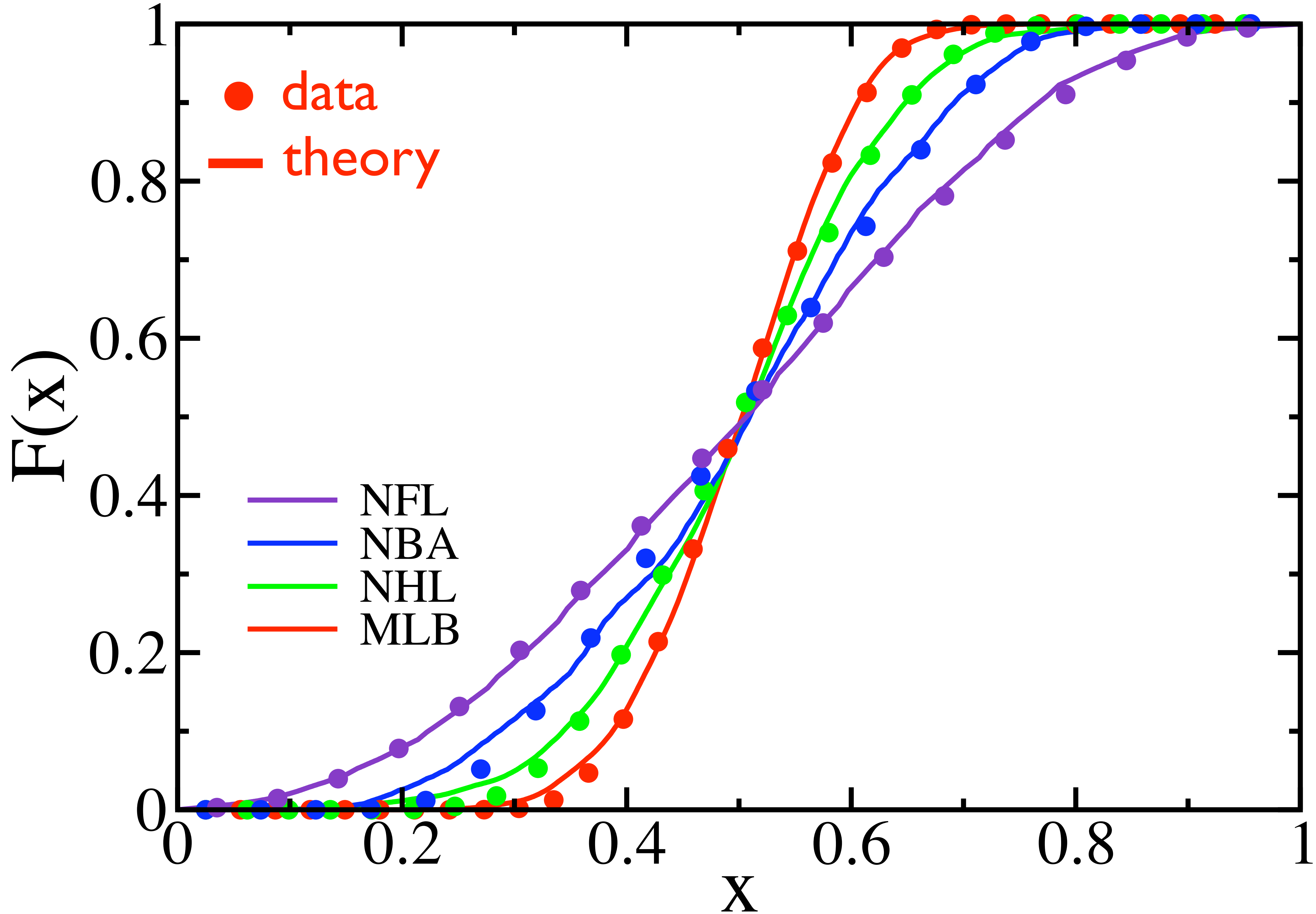
Data

- 300,000 Regular season games (all games)
- 5 Major sports leagues in US, England

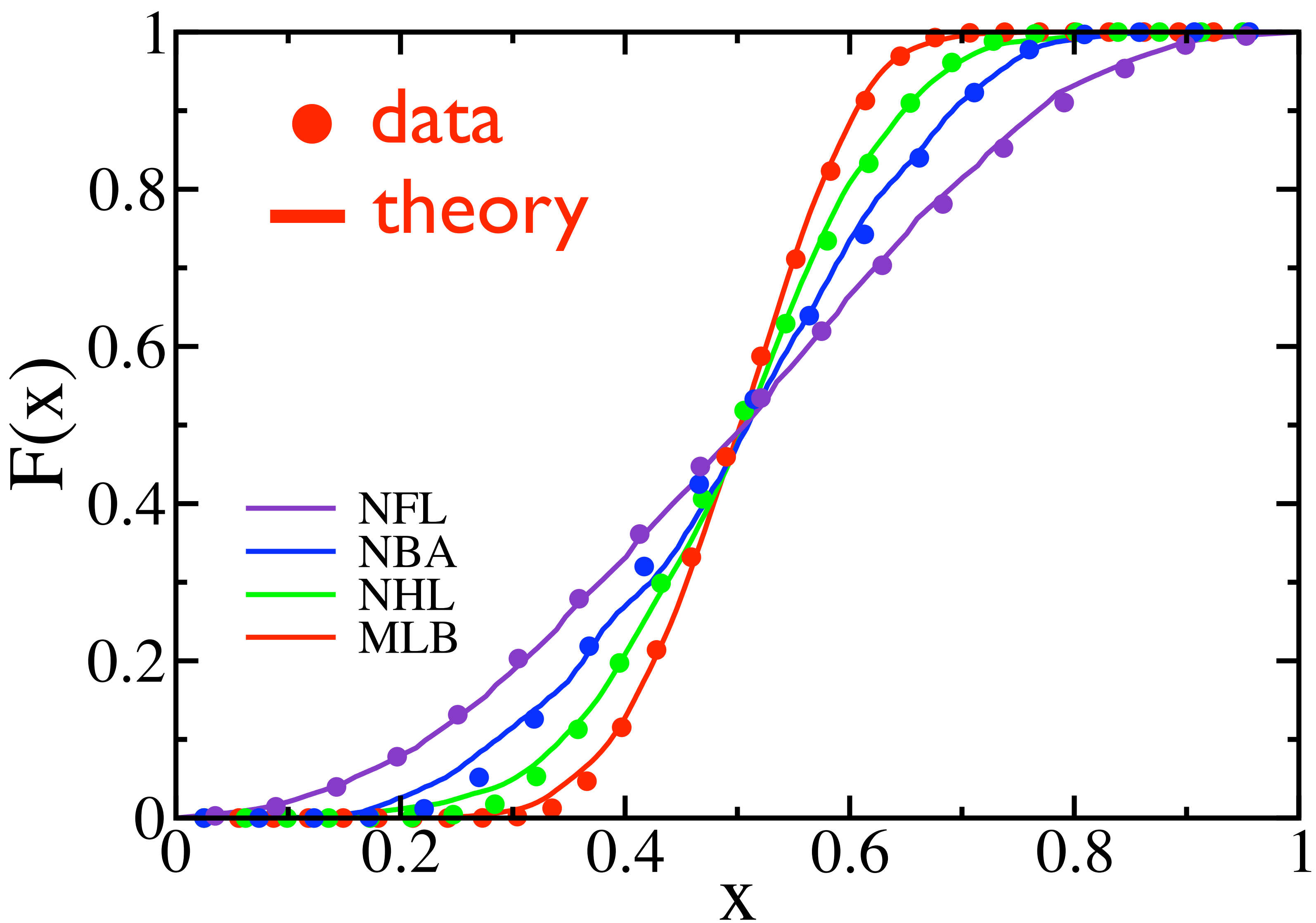
sport	league	full name	country	years	games
soccer	FA	Football Association	England	1888-2005	43,350
baseball	MLB	Major League Baseball	US	1901-2005	163,720
hockey	NHL	National Hockey League	US	1917-2005	39,563
basketball	NBA	National Basketball Association	US	1946-2005	43,254
american football	NFL	National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

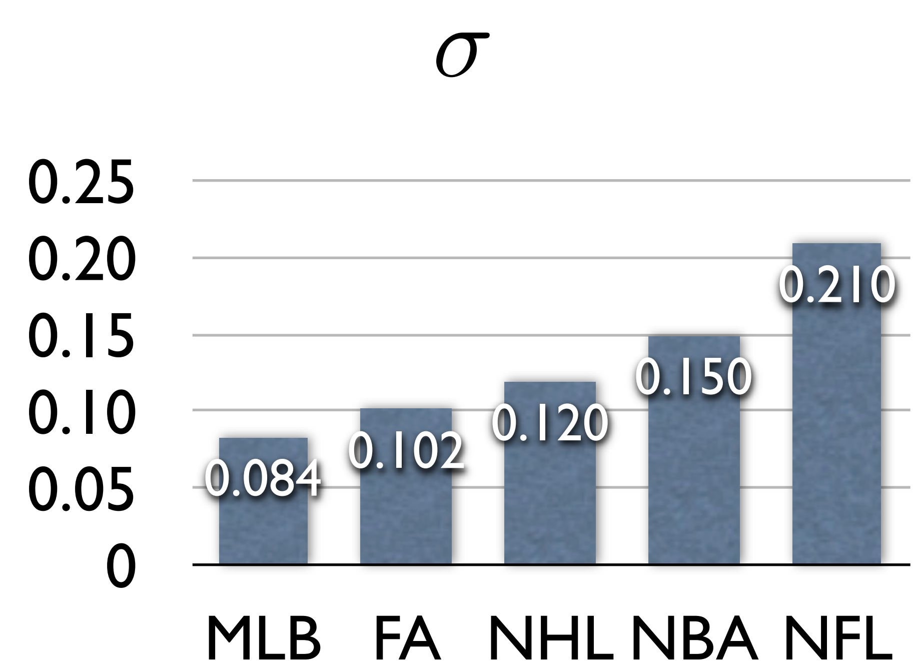
source: http://www.shrpsports.com/ http://www.the-english-football-archive.com/

Standard deviation in winning percentage



Standard deviation in winning percentage





- Baseball most competitive?
- American football least competitive?

Distribution of winning percentage clearly distinguishes sports

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

$$q = \begin{cases} 1/2 & \text{random} \\ 1 & \text{deterministic} \end{cases}$$

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

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

Rate equation approach

- Probability distribution functions

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

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

- 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 $G_0 = 0 \qquad G_\infty = 1$ 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

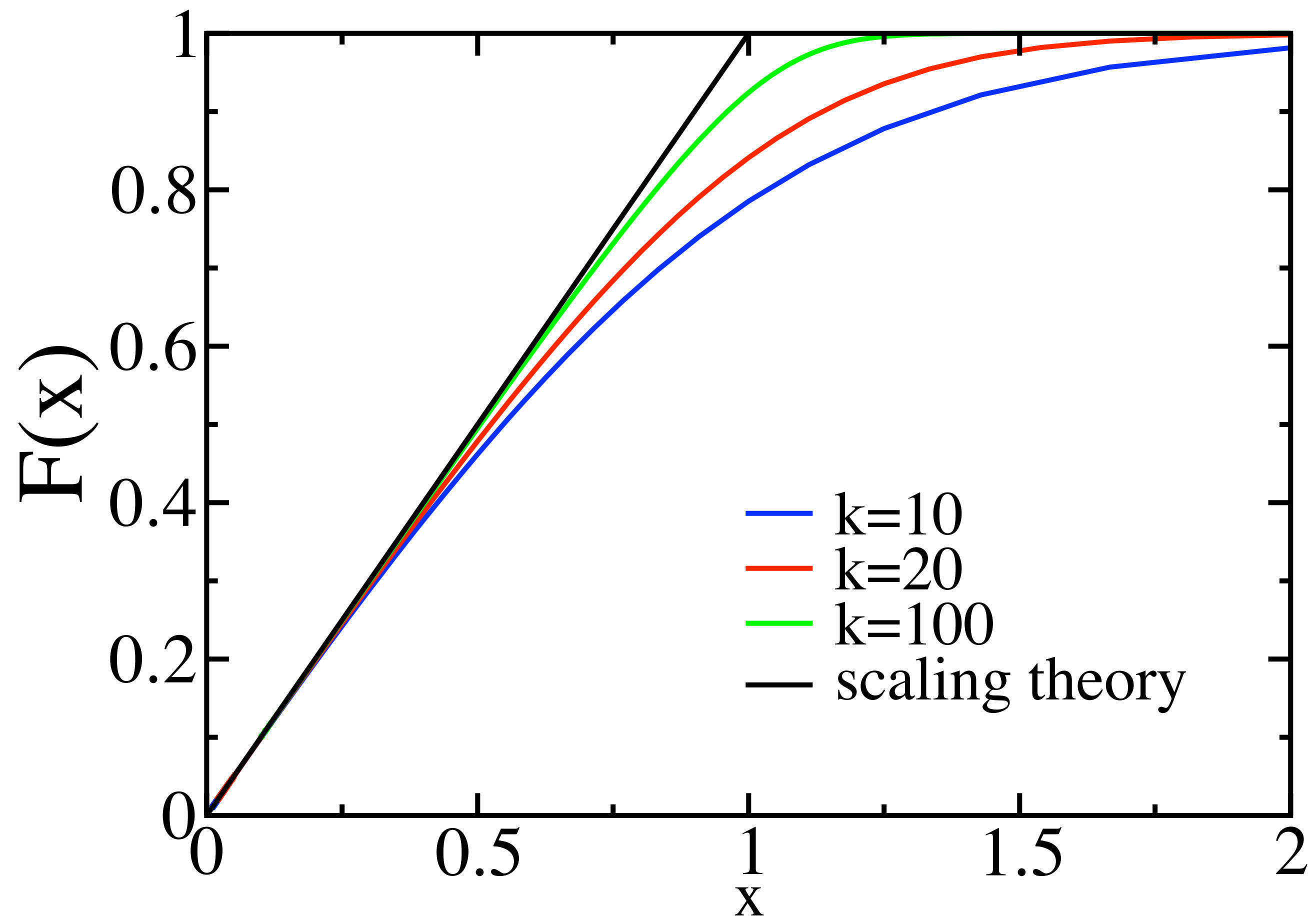
$$G_k \to \frac{k+1}{t}$$

- 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

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

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

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

- 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 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

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

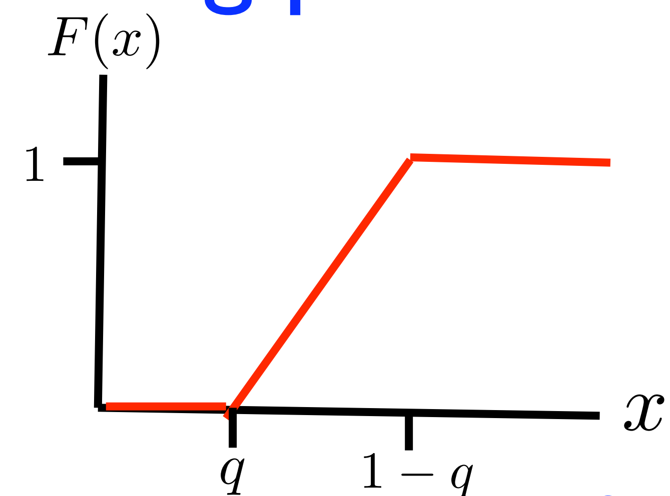
- Scaling equation

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

Scaling solution

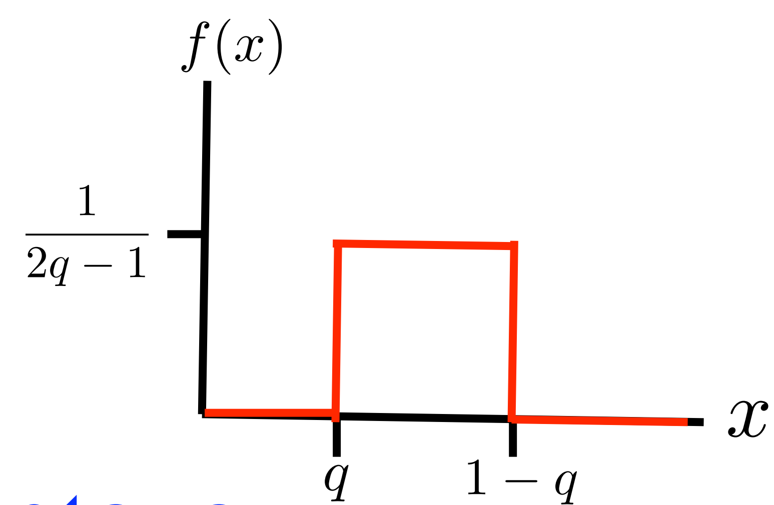
- Stationary distribution of winning percentage

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



- Distribution of winning percentage is uniform

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

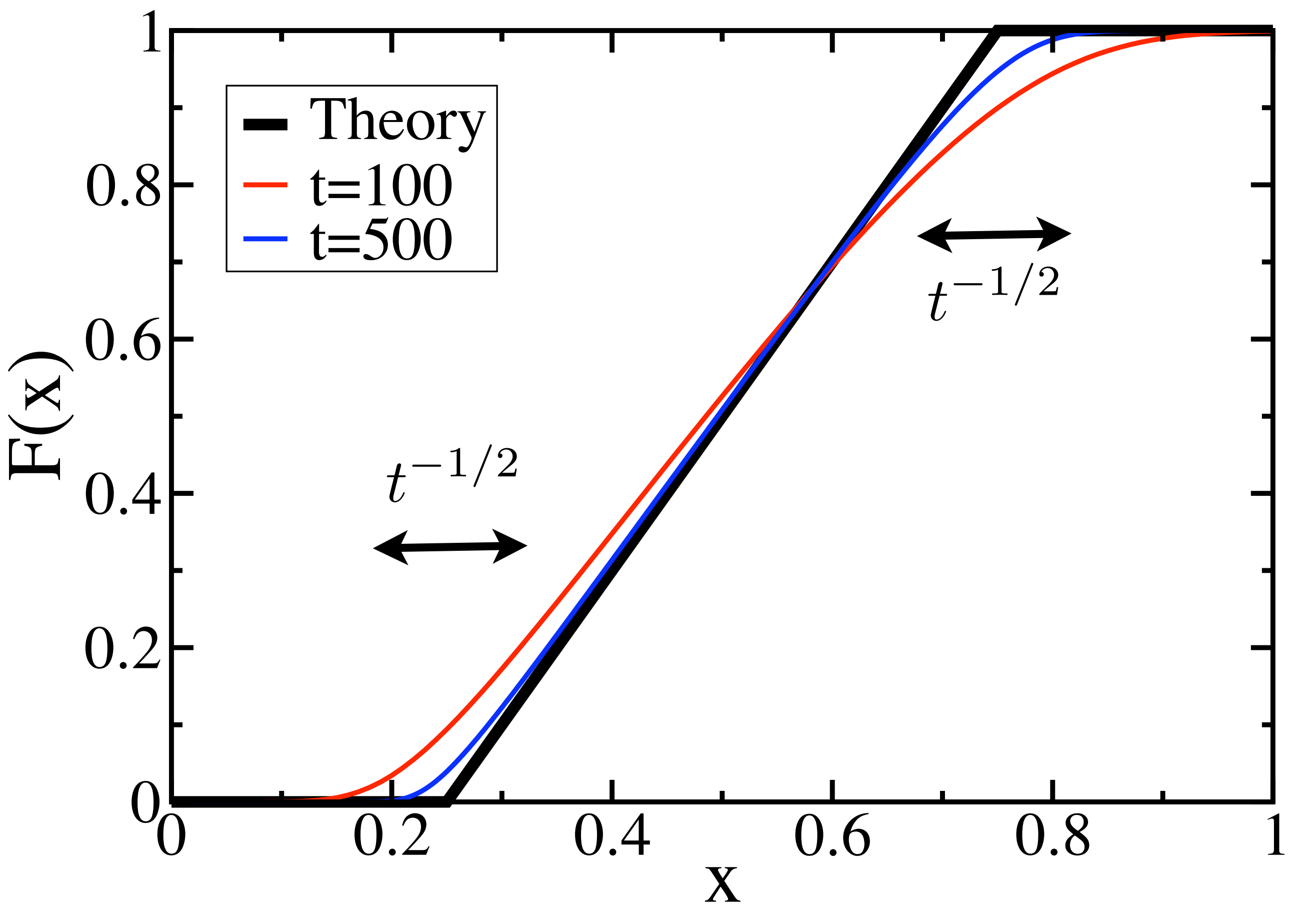


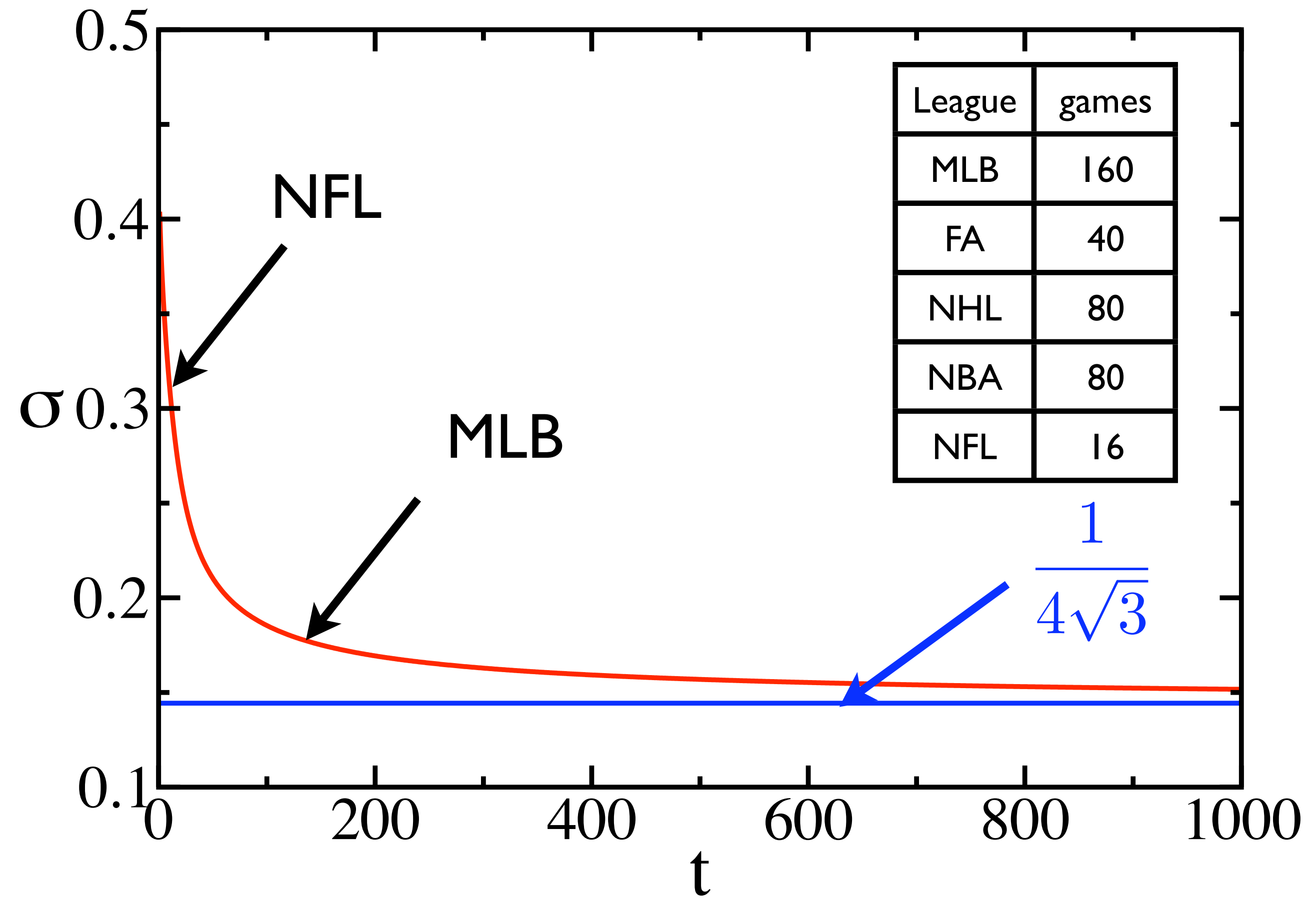
- Variance in winning percentage

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

Approach to scaling

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





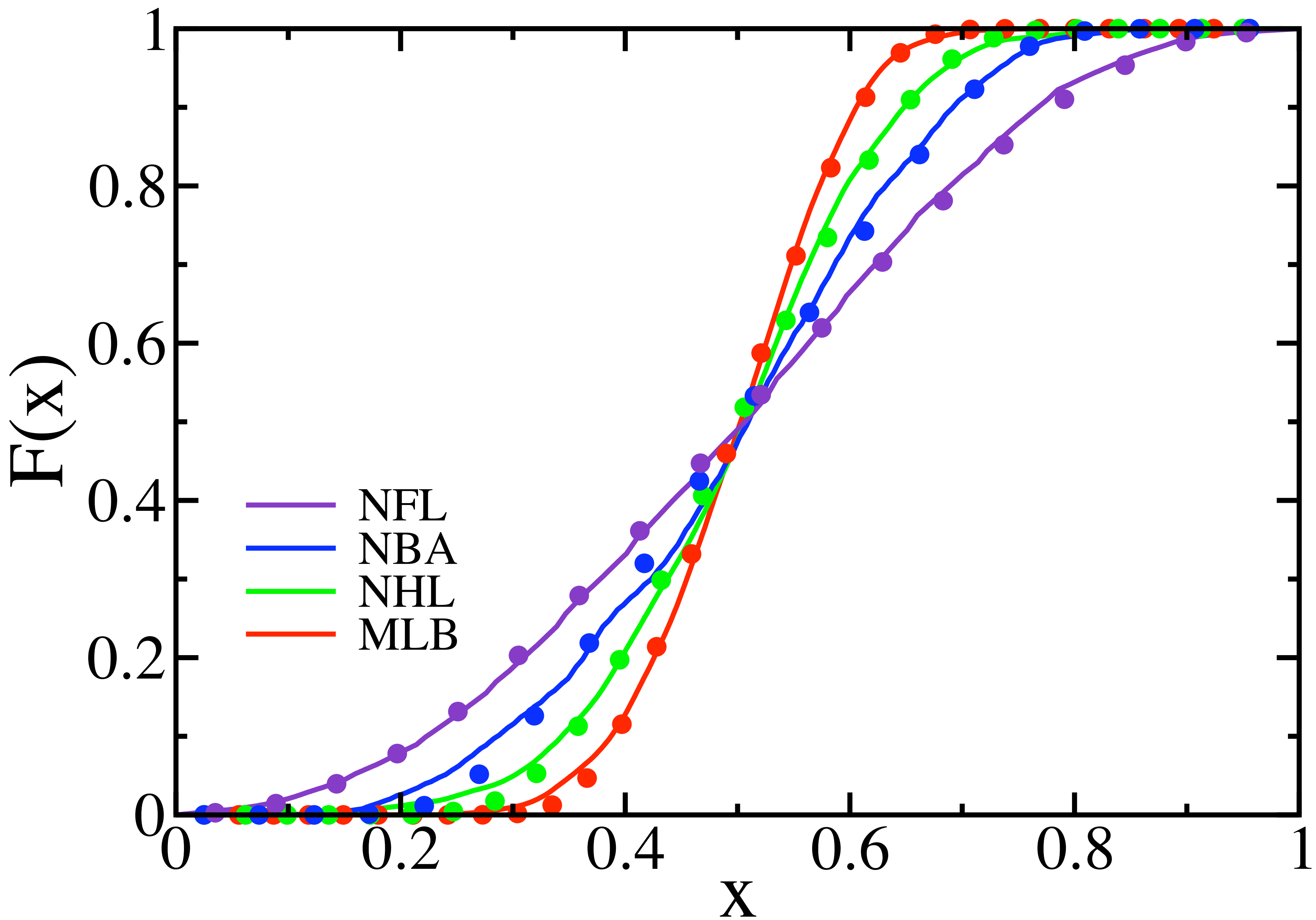
League	games
MLB	160
FA	40
NHL	80
NBA	80
NFL	16

- 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

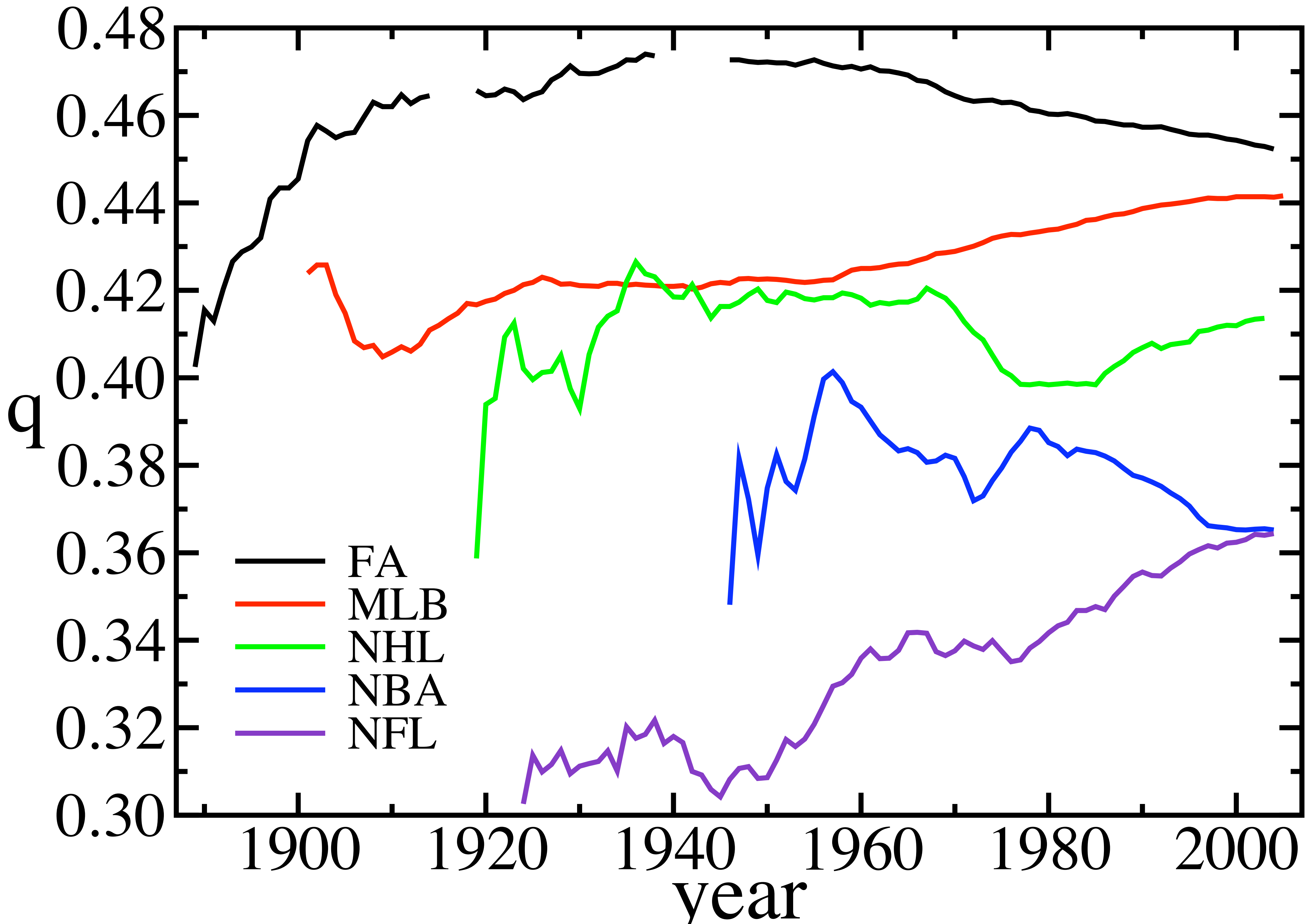
The upset frequency

- Upset frequency as a measure of predictability

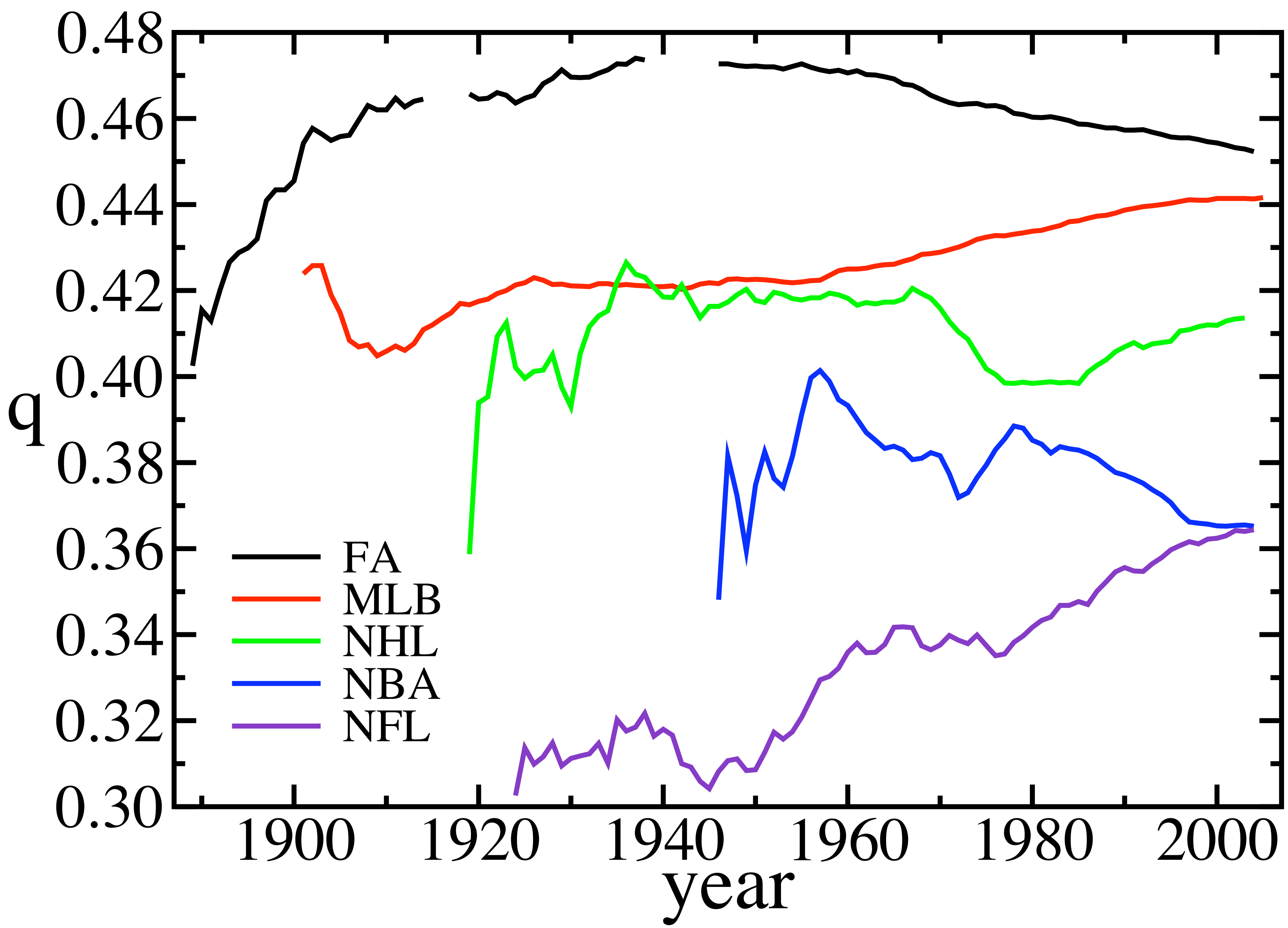
$$q = \frac{\text{Number of upsets}}{\text{Number of games}}$$

- 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



The upset frequency



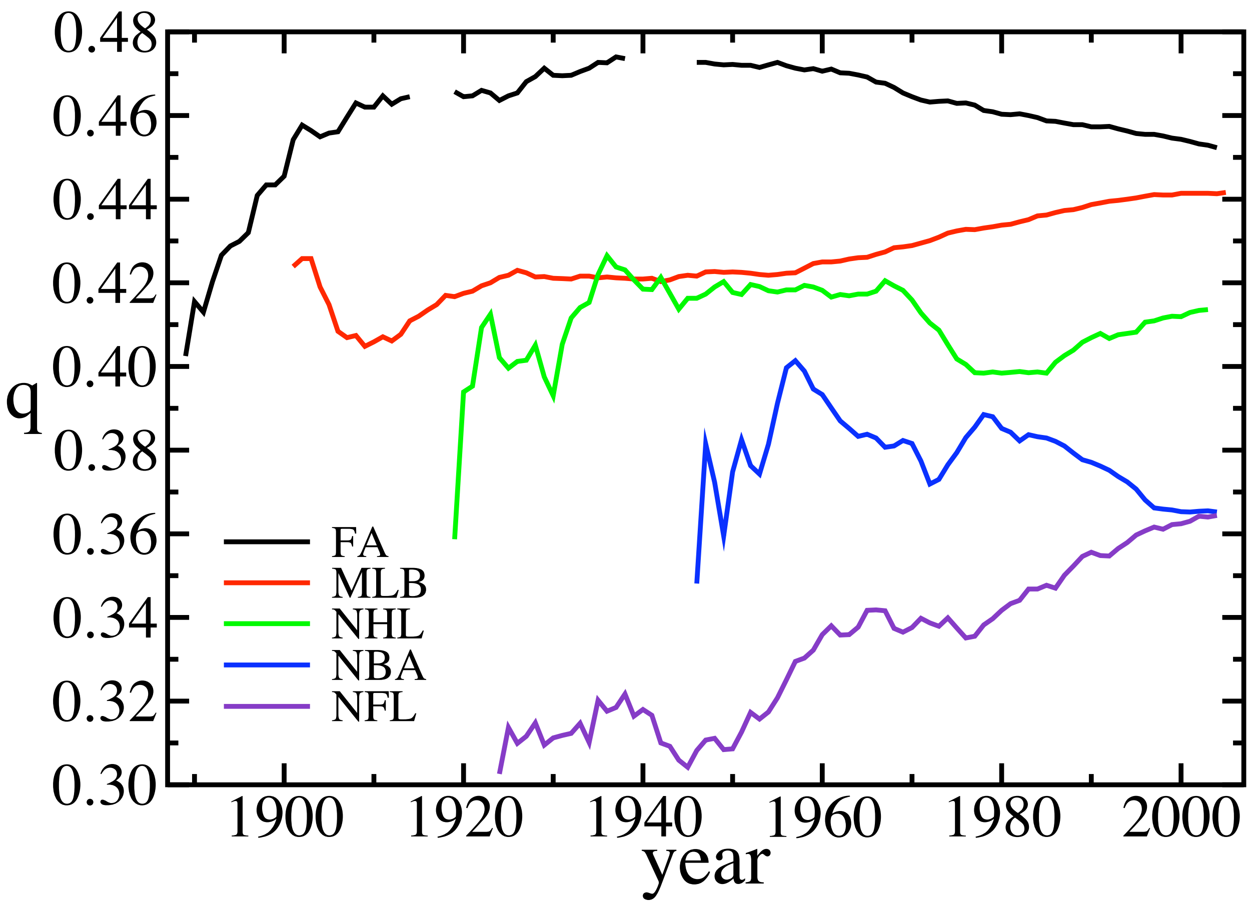
League	q	q_{model}
FA	**0.452**	0.459
MLB	**0.441**	0.413
NHL	**0.414**	0.383
NBA	**0.365**	0.316
NFL	**0.364**	0.309

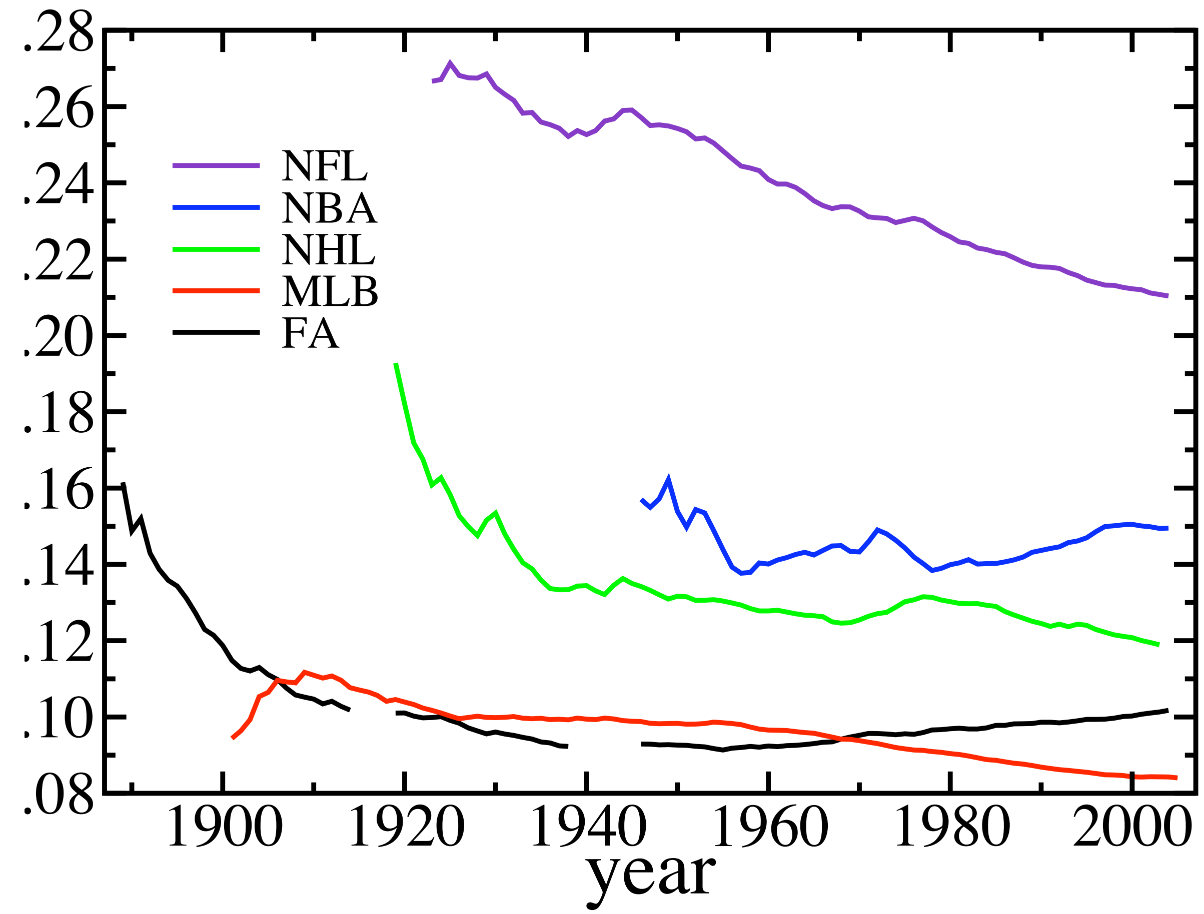
q differentiates the different sport leagues!

Football, baseball most competitive

Basketball, American football least competitive

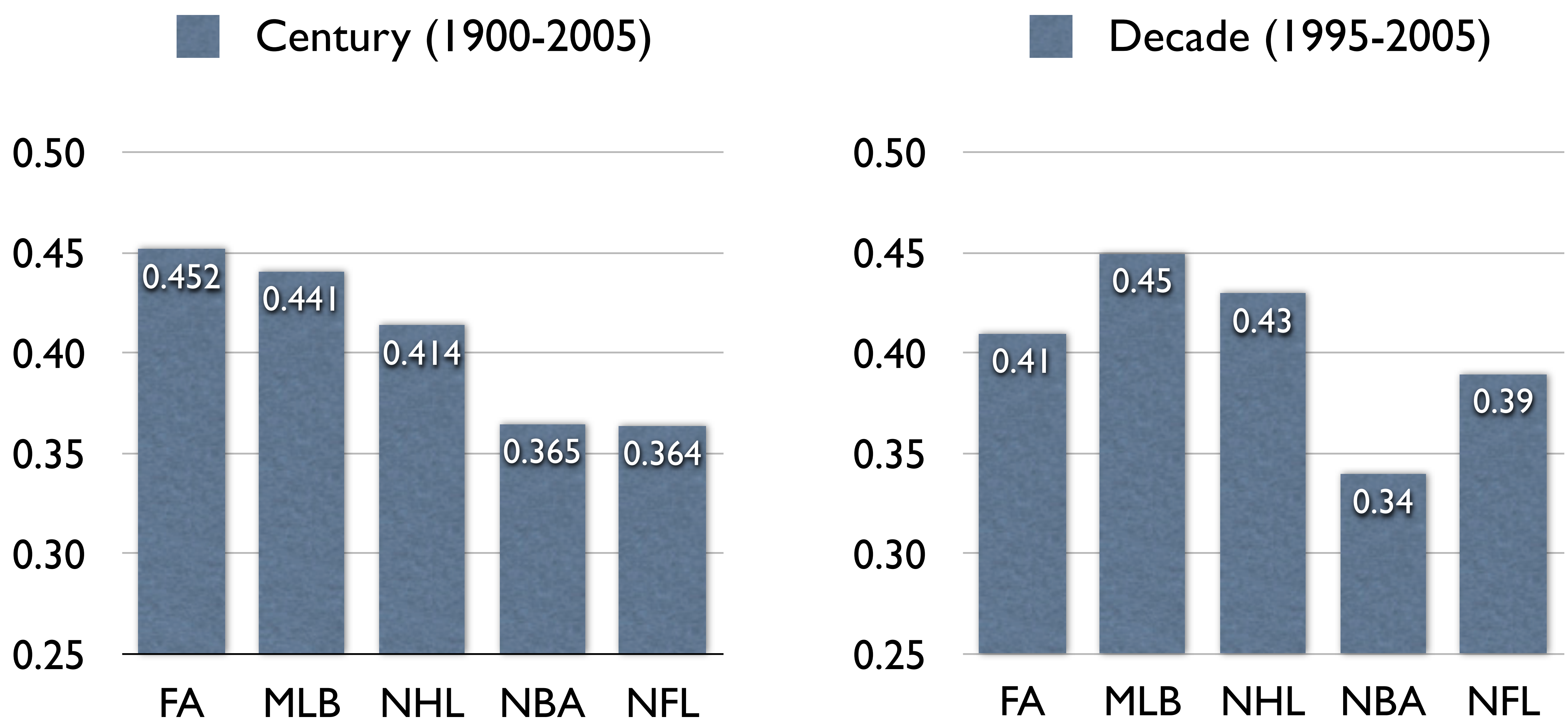
Evolution with time





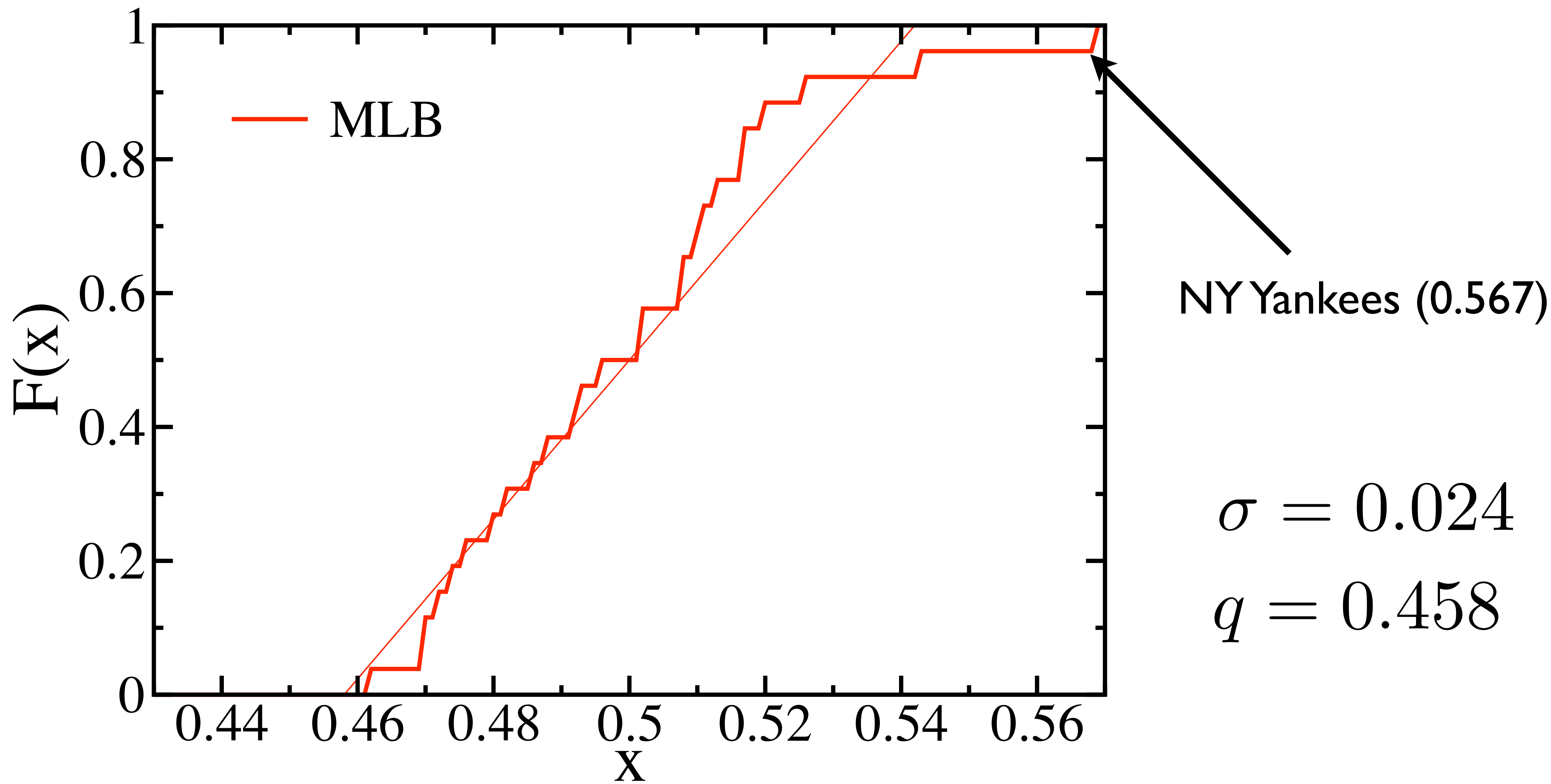
- Parity, predictability mirror each other $\sigma = \frac{1/2 - q}{\sqrt{3}}$
- American football, baseball increasing competitiveness
- Football decreasing competitiveness (past 60 years)

Century versus Decade



Football-American Football gap narrows from 9% to 2%!

All-time team records



$$\sigma = 0.024$$
$$q = 0.458$$

- Provides the longest possible record (t~13000)
- Close to a linear function

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

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

- 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

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

Social structures

1. Middle class

Agents advance at different rates

2. Middle+lower class

Some agents advance at different rates

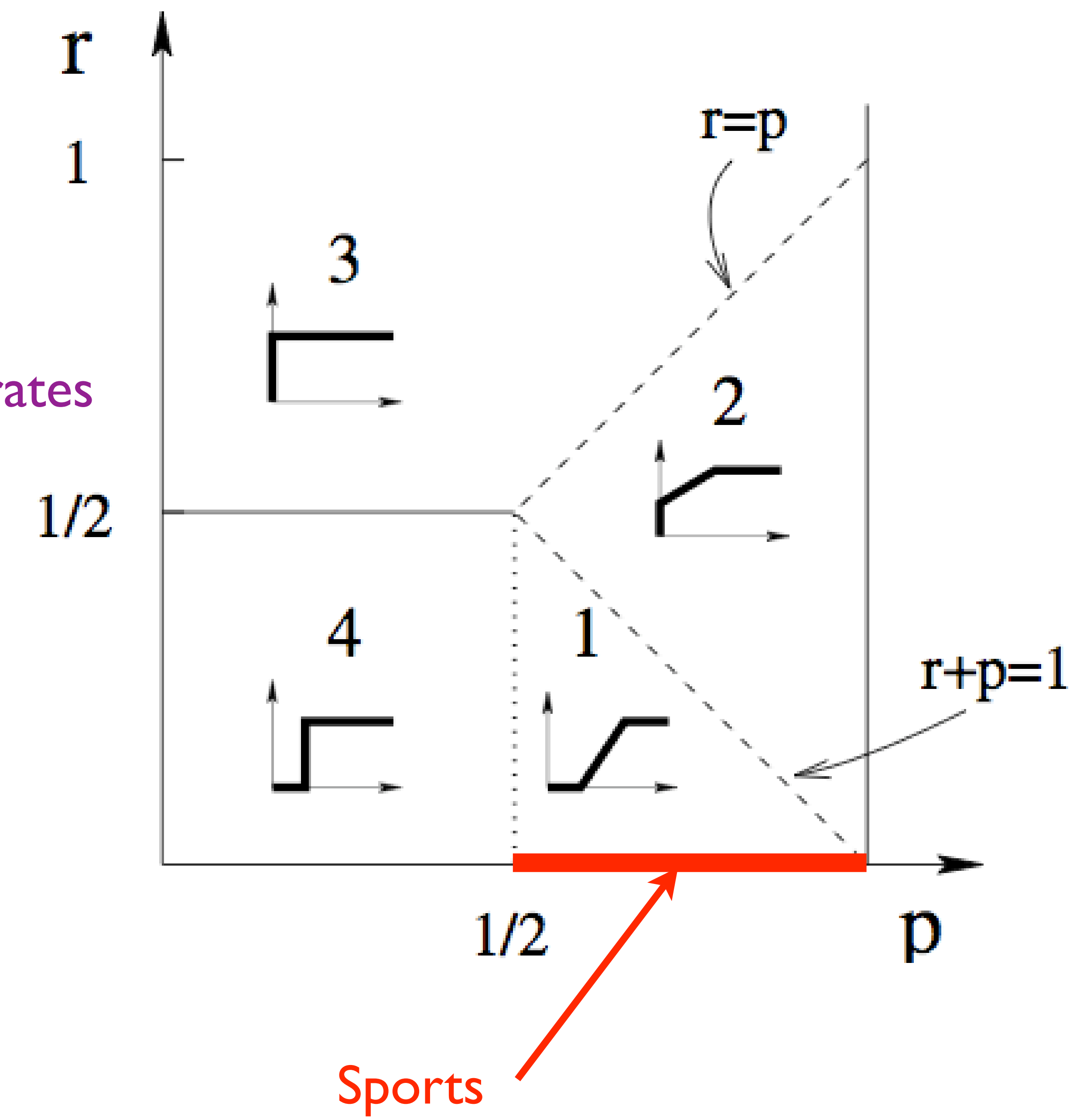
Some agents do not advance

3. Lower class

Agents do not advance

4. Egaliterian class

All agents advance at equal rates



Bonabeau 96

Publications

- Parity and Predictability of Competitions
  E. Ben-Naim, F. Vazquez, S. Redner
  J. Quant. Anal. in Sports, submitted (2006)
- 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
  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)

“I do not make predictions,

especially not about the future.”

Yogi Bera