Scoring dynamics in professional sports: tempo, balance, predictability

Sears Merritt, Aaron Clauset

Why study sports competitions?

Ideal environment to study fundamental properties of competition.

- Level playing field
- Clear and enforceable rules
- Copious amounts of detailed, longitudinal data

Outline

- Model of competition
- Data set
- Analysis of timing, balance, points
- Simulation
- Prediction

Model of competition

Tempo

Probability of event occurring at time t

$\Pr(\text{event})(t)$

Balance

Probability of winning an event

$\Pr(S_r \text{ wins})$

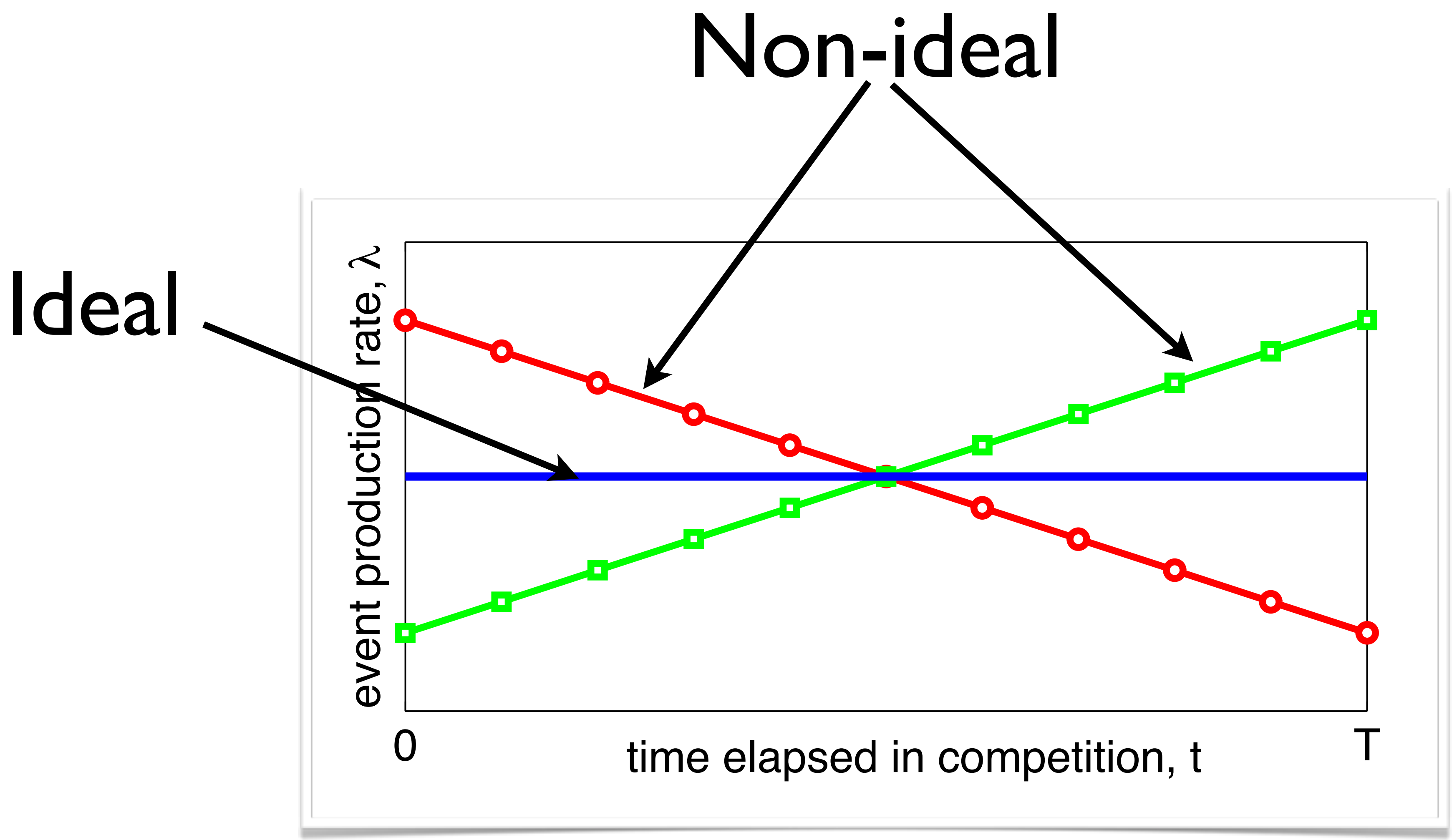
Points

Probability of event being worth i points

$\Pr(\text{points} = i)$

$$\Pr(\Delta S_r = i)(t) = \Pr(\text{event})(t) \, \Pr(S_r \text{ wins}) \, \Pr(\text{points} = i)$$

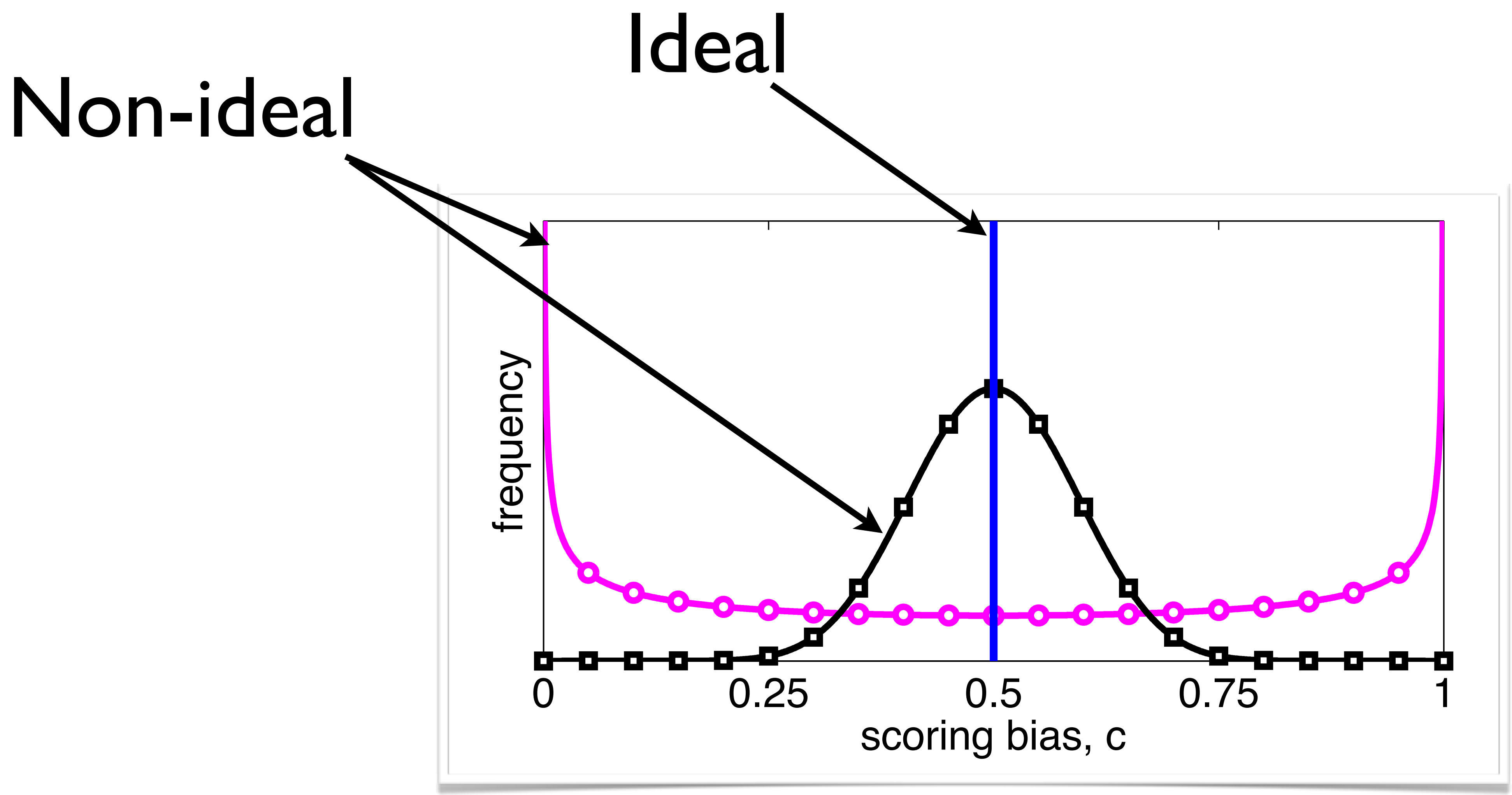
Tempo



Ideal $\sim \mathrm{Poisson}(\lambda)$

Non-ideal $\lambda(t) = \lambda_0 + \alpha(t)$

Balance



Ideal $\sim \text{Bernoulli}(c = 1/2)$

Non-ideal $\Pr(c) = \text{Beta}(\beta, \beta)$

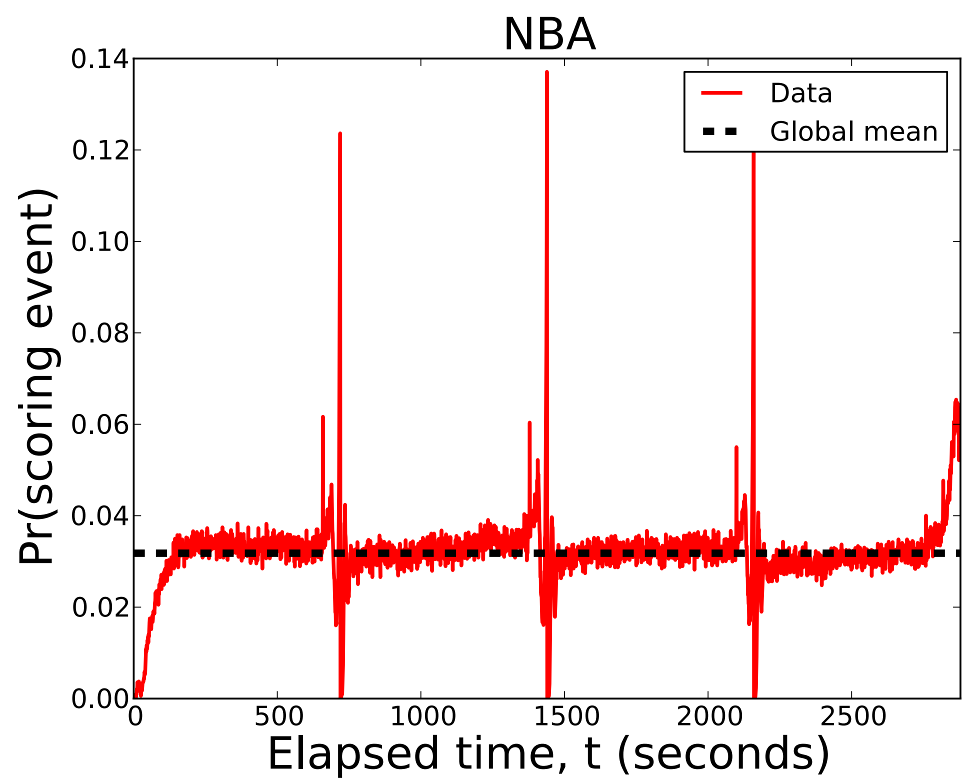
Scoring event data

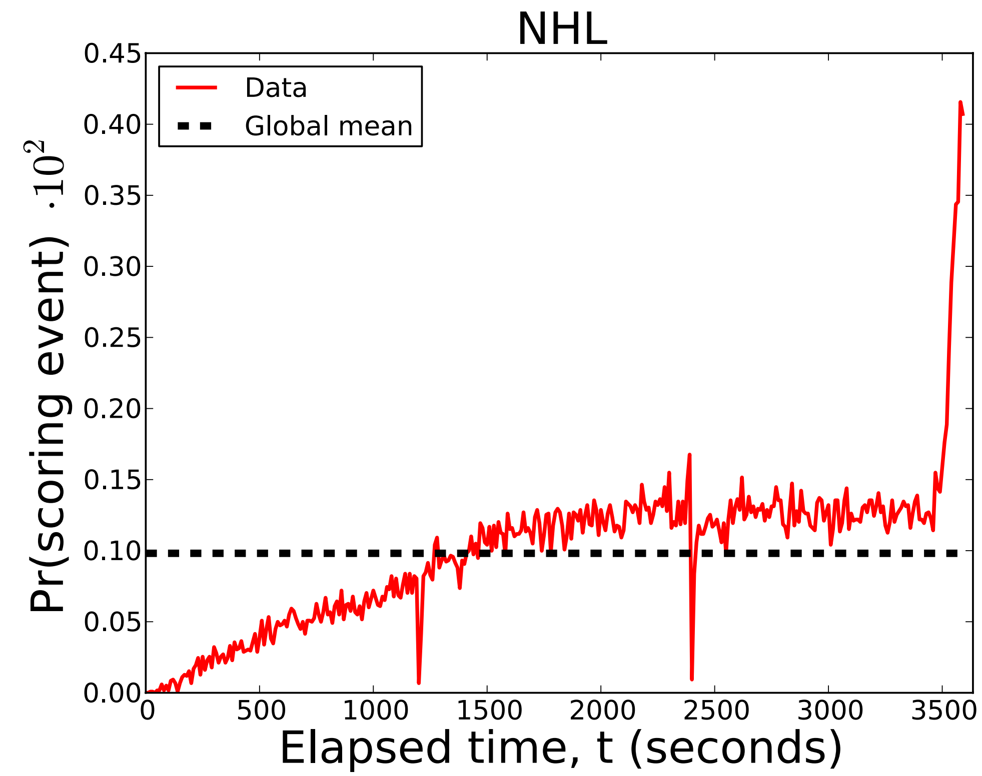
sport	abbr.	time	competitions	scoring events	total teams
Pro. football	NFL	2000-2009	2,654	19,814	31
Col. football	CFB	2000-2009	14,588	120,829	486
Pro. hockey	NHL	2000-2009	11,744	47,539	29
Pro. basketball	NBA	2002-2010	11,813	1,091,719	31

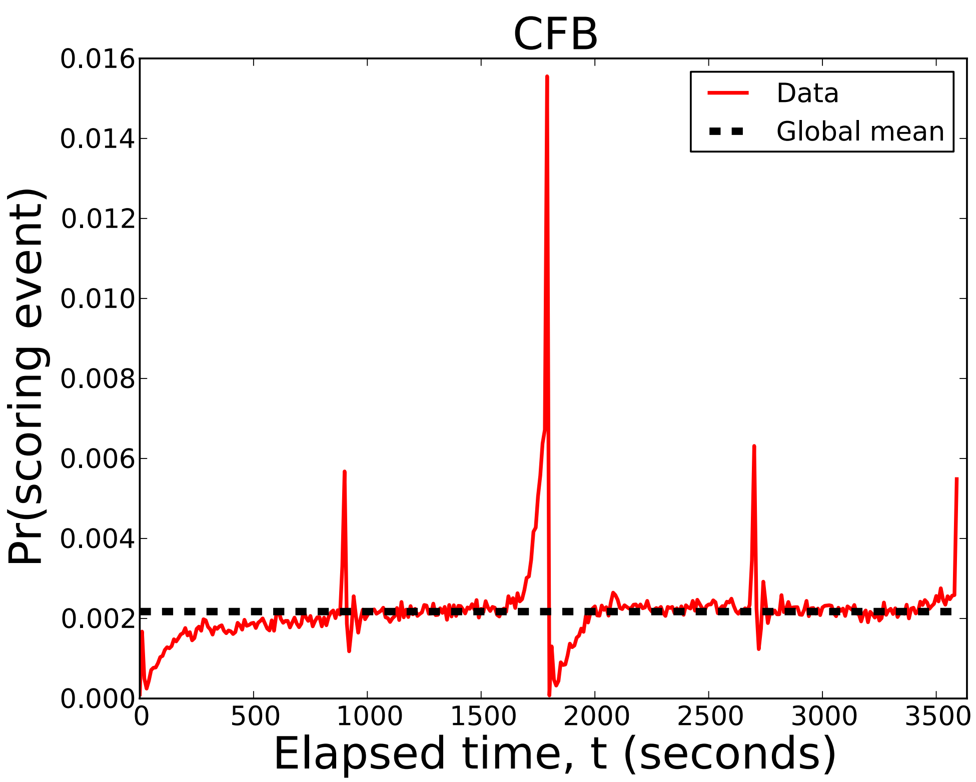
Scoring event: (team, game clock, point value)

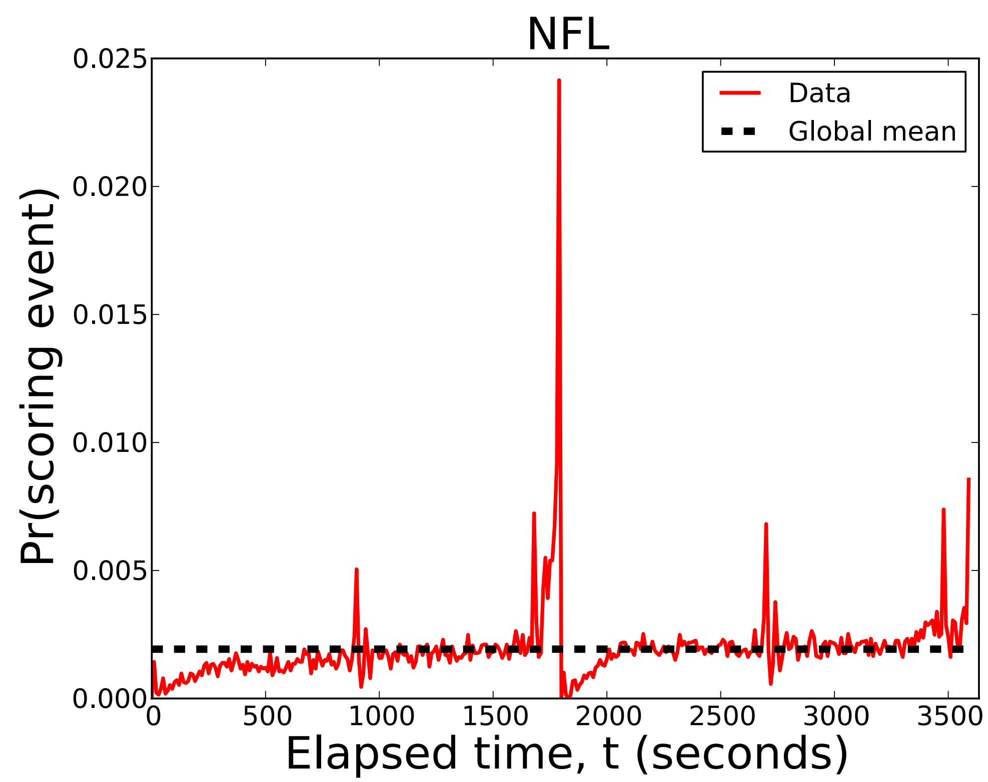
to nearest second

Tempo

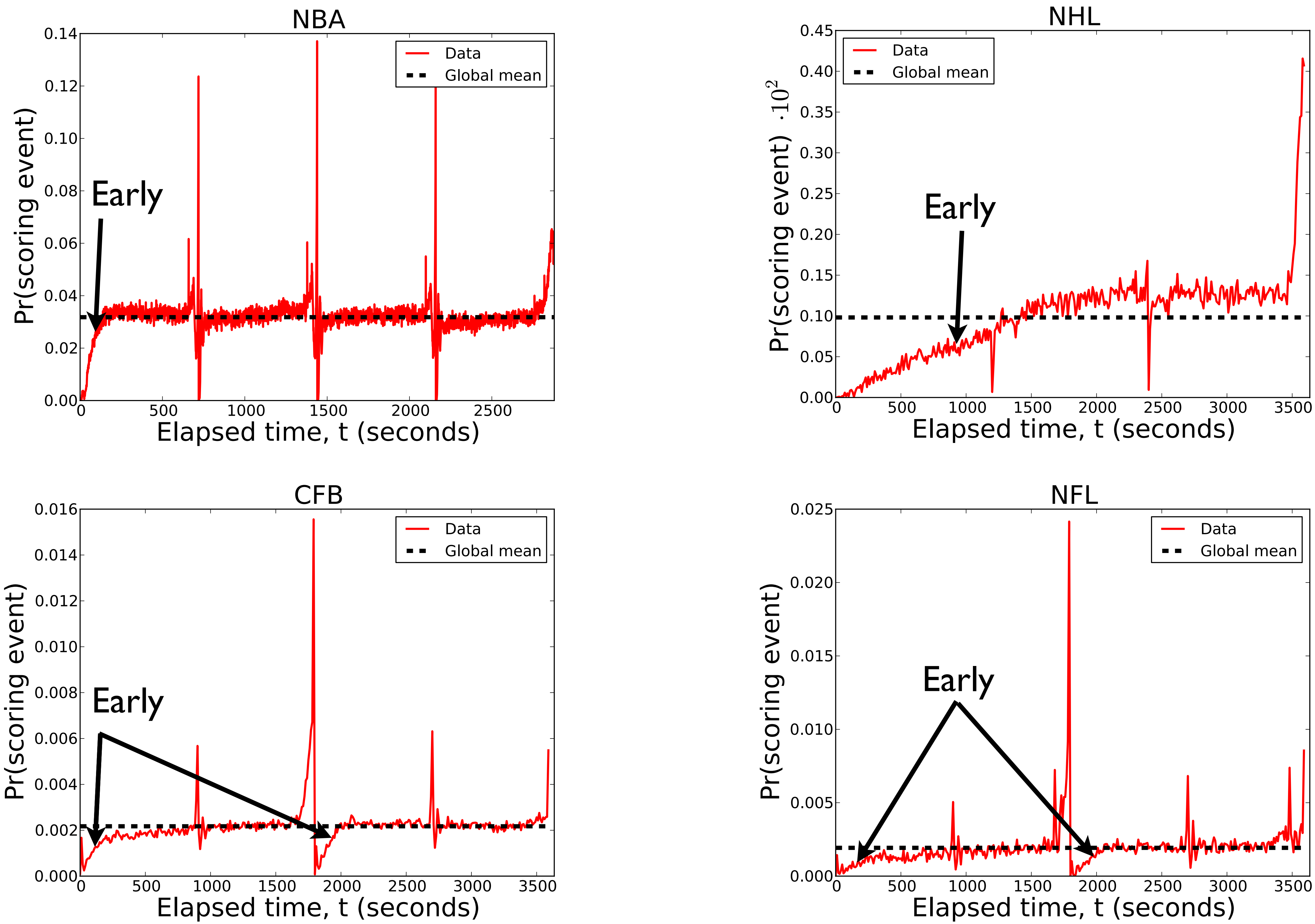

NBA
Data
Global mean
Pr(scoring event)
Elapsed time, t (seconds)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0
500
1000
1500
2000
2500


NHL
Data
Global mean
Pr(scoring event) $\cdot 10^2$
Elapsed time, t (seconds)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0
500
1000
1500
2000
2500
3000
3500

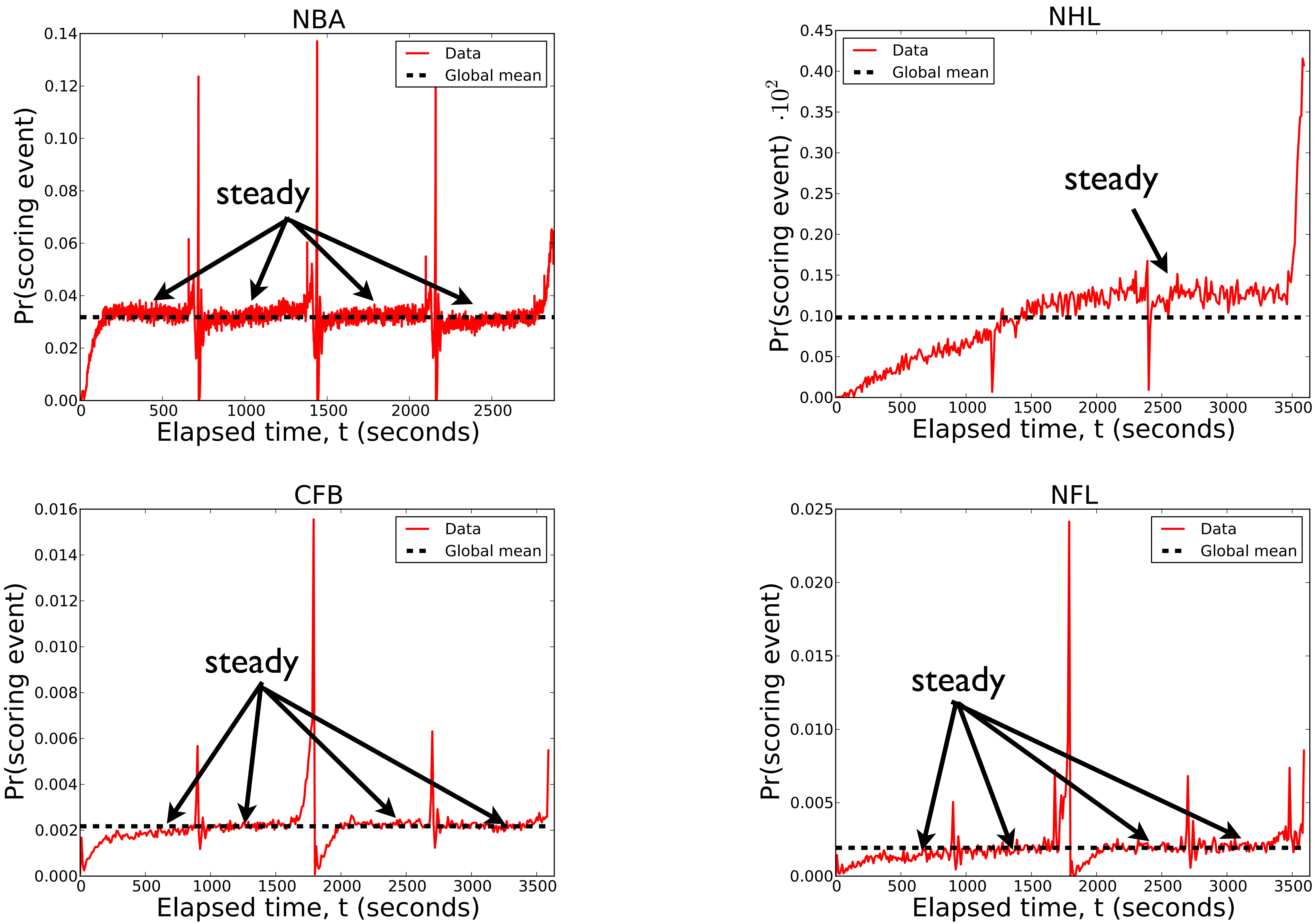

CFB
Data
Global mean
Pr(scoring event)
Elapsed time, t (seconds)
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
0.016
0
500
1000
1500
2000
2500
3000
3500


NFL
Data
Global mean
Pr(scoring event)
Elapsed time, t (seconds)
0.000
0.005
0.010
0.015
0.020
0.025
0
500
1000
1500
2000
2500
3000
3500

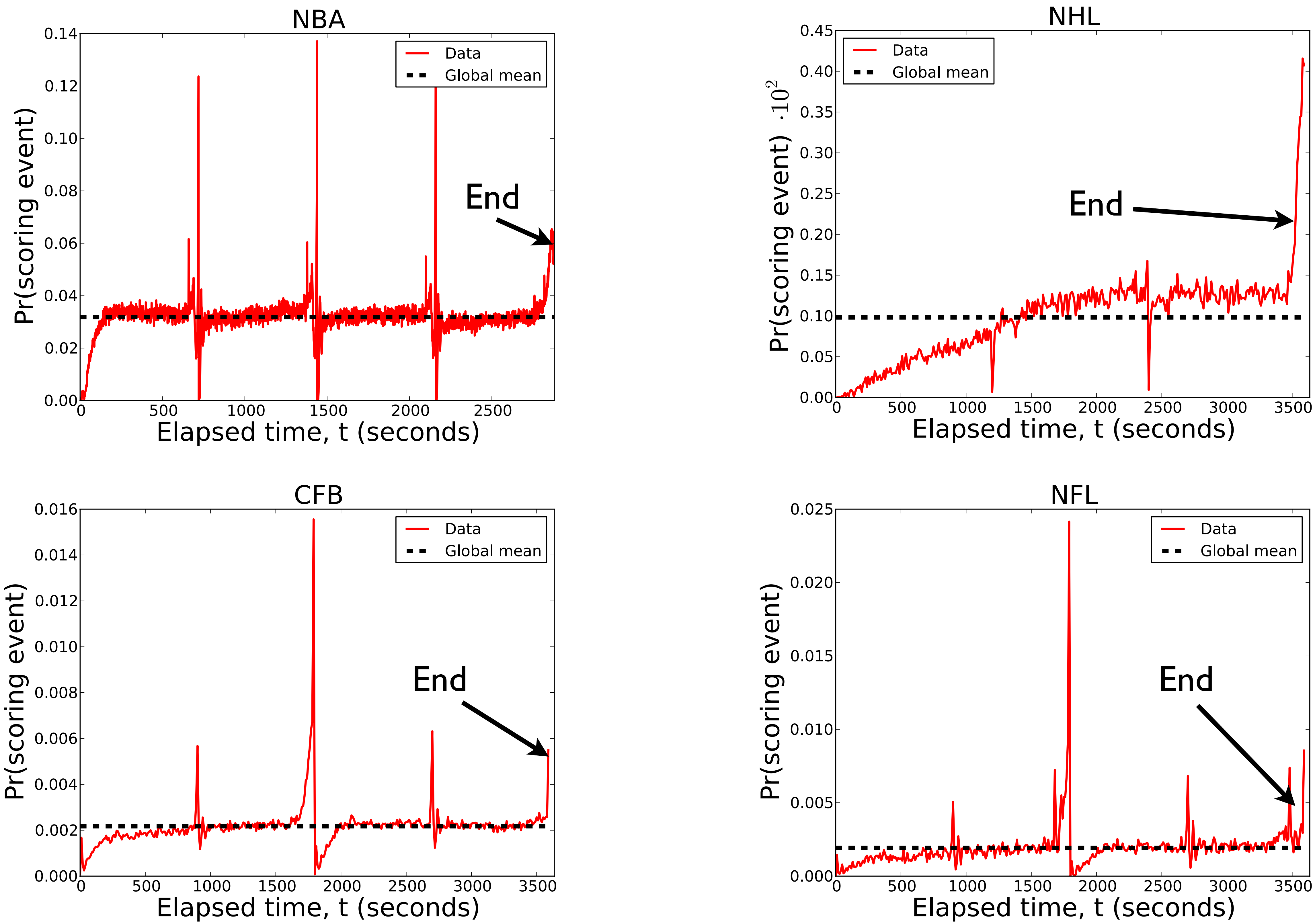
Tempo: early phase



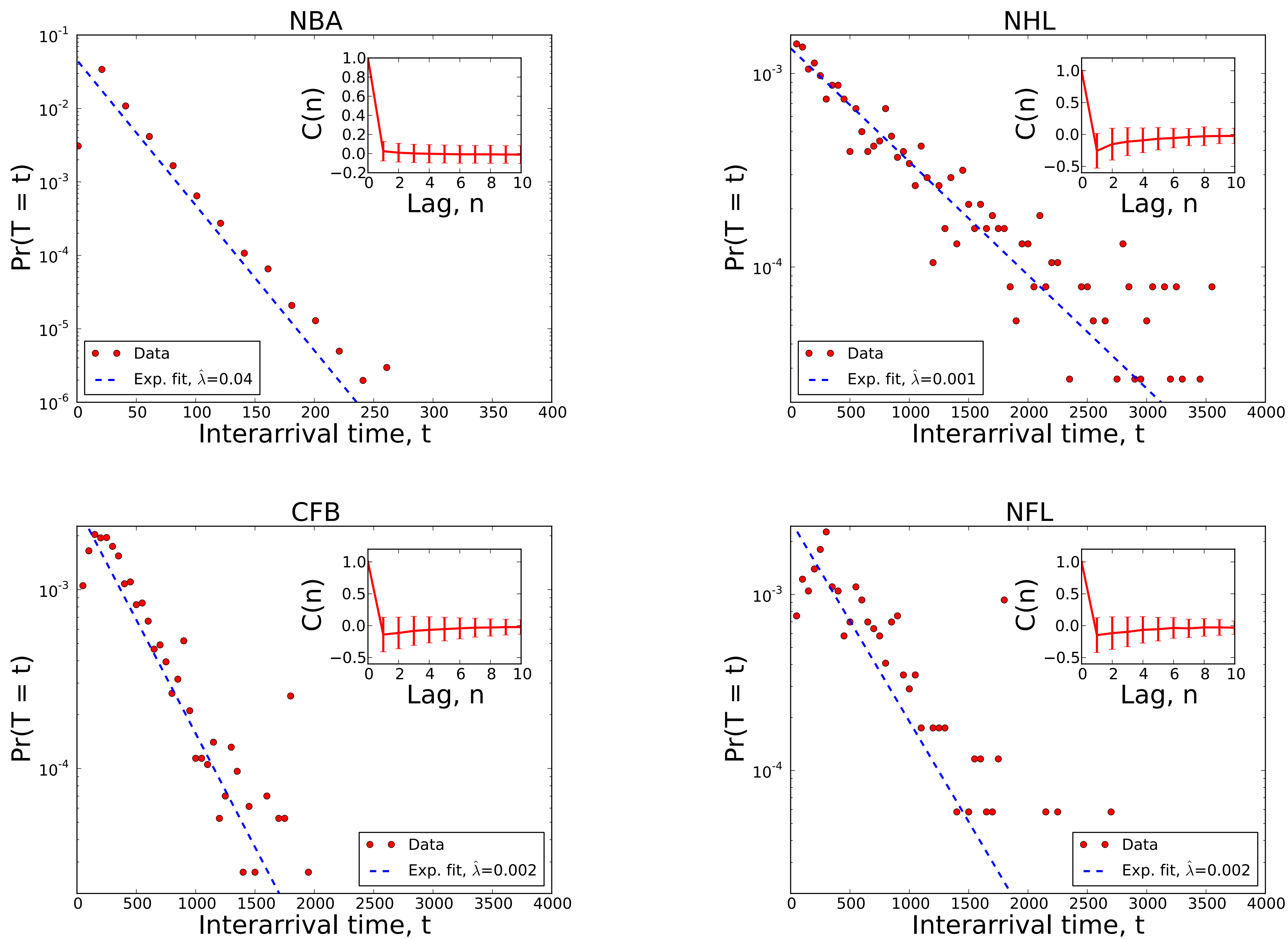
Tempo: steady phase



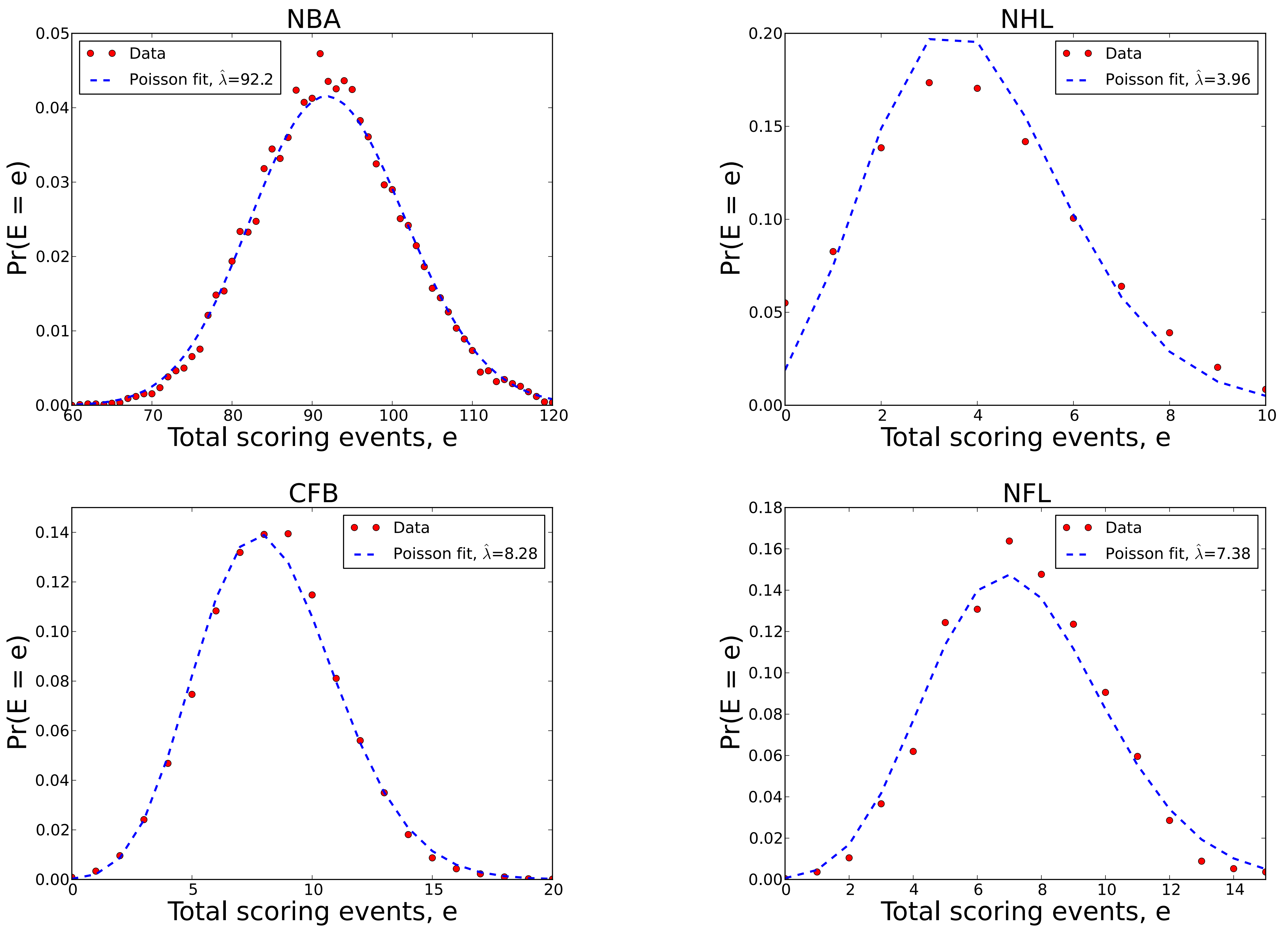
Tempo: end phase



Tempo: inter-arrivals, correlation



Timing - cumulative events



Balance

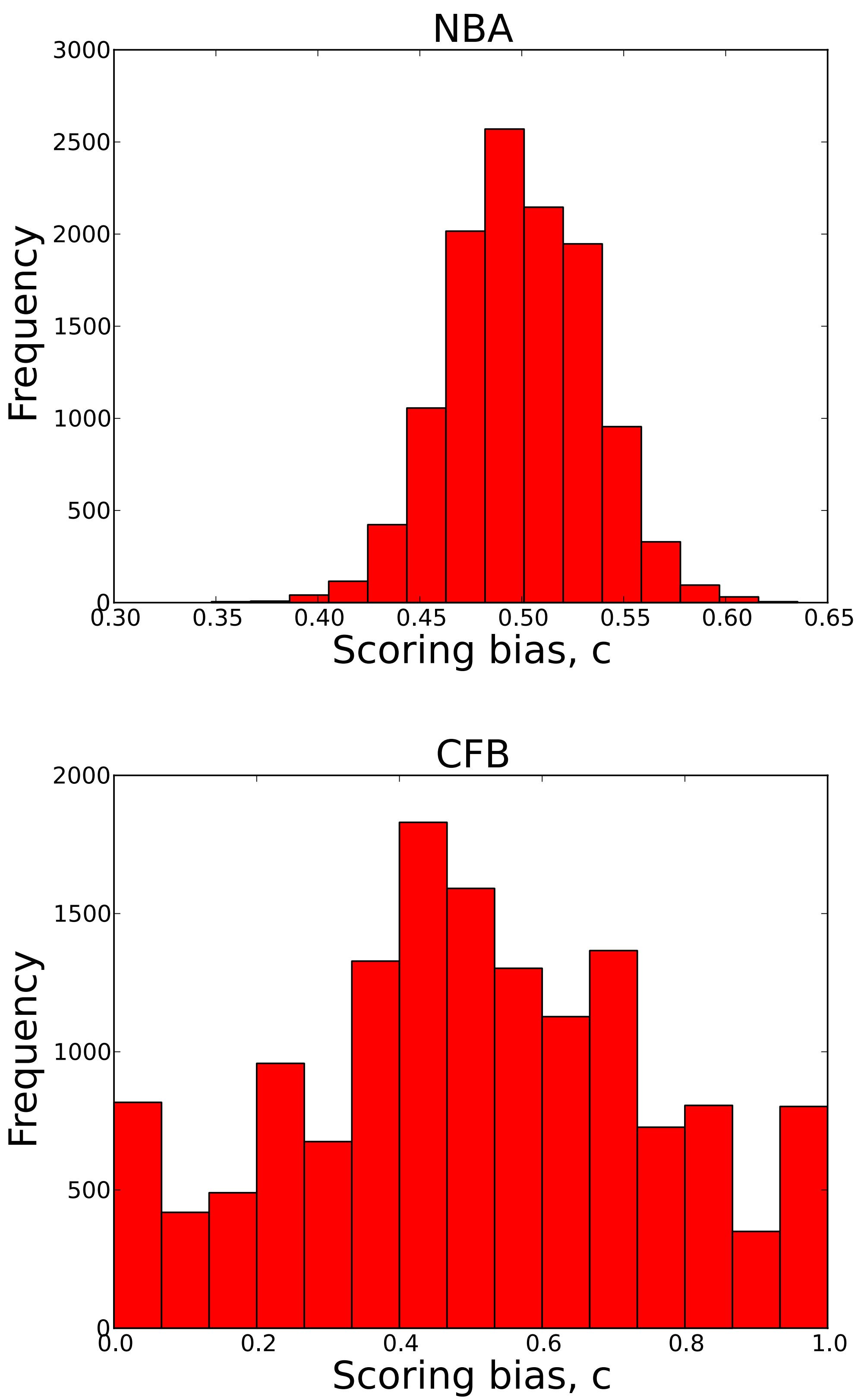
Independent scoring events

$$\Pr(S_r) = \binom{S_r + S_b}{S_r} c^{S_r} (1 - c)^{S_b}$$

Maximum likelihood estimator

$$\hat{c} = \frac{S_r}{S_r + S_b}$$

Balance



NBA
Frequency
Scoring bias, c
NHL
Frequency
Scoring bias, c
CFB
Frequency
Scoring bias, c
NFL
Frequency
Scoring bias, c

Balance

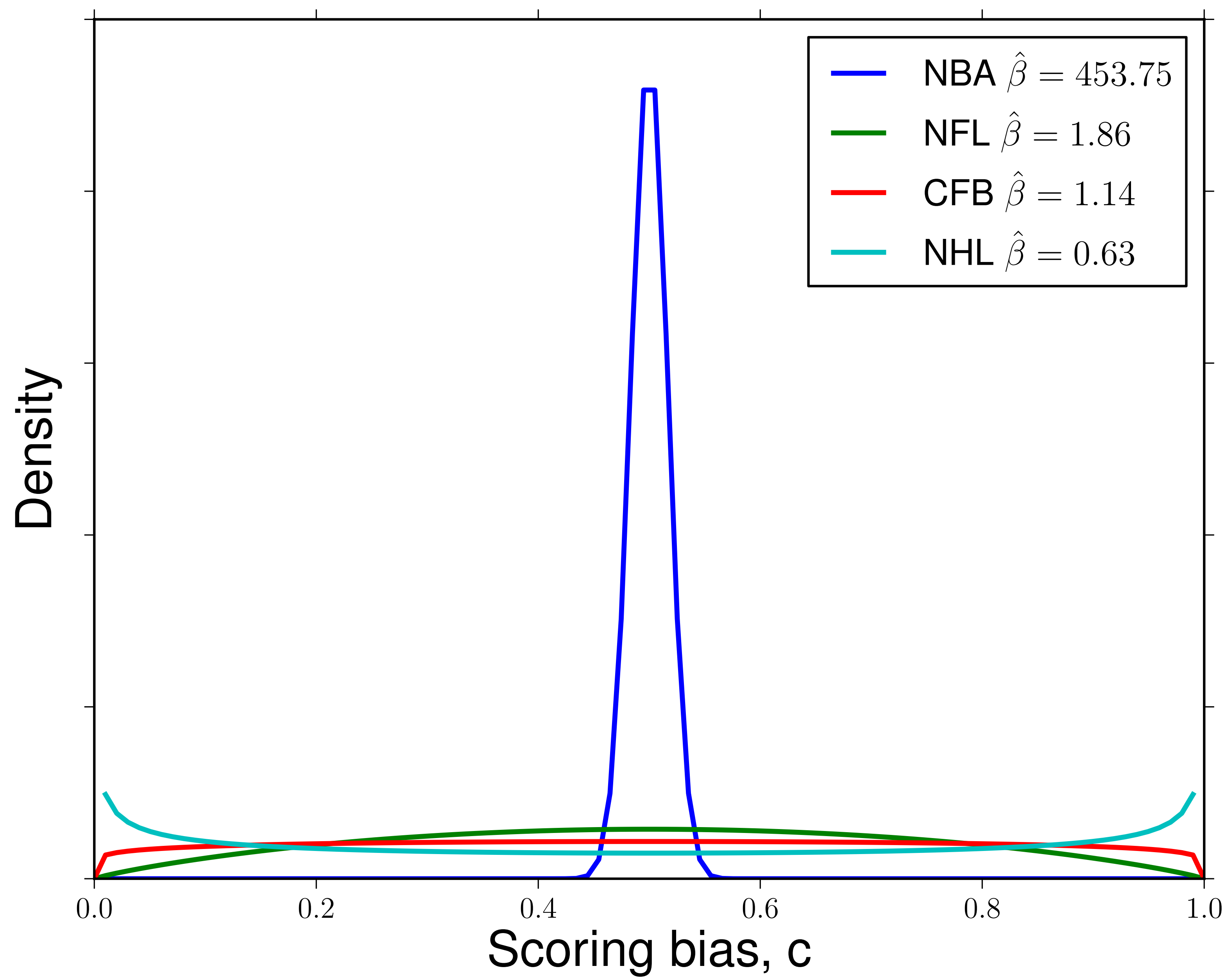
$$\mathcal{L} = \prod_{k=1}^{N} \Pr(S_{r_k}, S_{b_k} | c) \Pr(c)$$

$$\mathcal{L} = \prod_{k=1}^{N} c^{S_{r_k}} (1-c)^{S_{b_k}} \frac{c^{\beta-1}(1-c)^{\beta-1}}{\mathrm{B}(\beta,\beta)}$$

$$\ln \mathcal{L} = \sum_{k=1}^{N} \ln[\mathrm{B}(S_{r_k}+\beta, S_{b_k}+\beta)] - \ln[\mathrm{B}(\beta,\beta)]$$

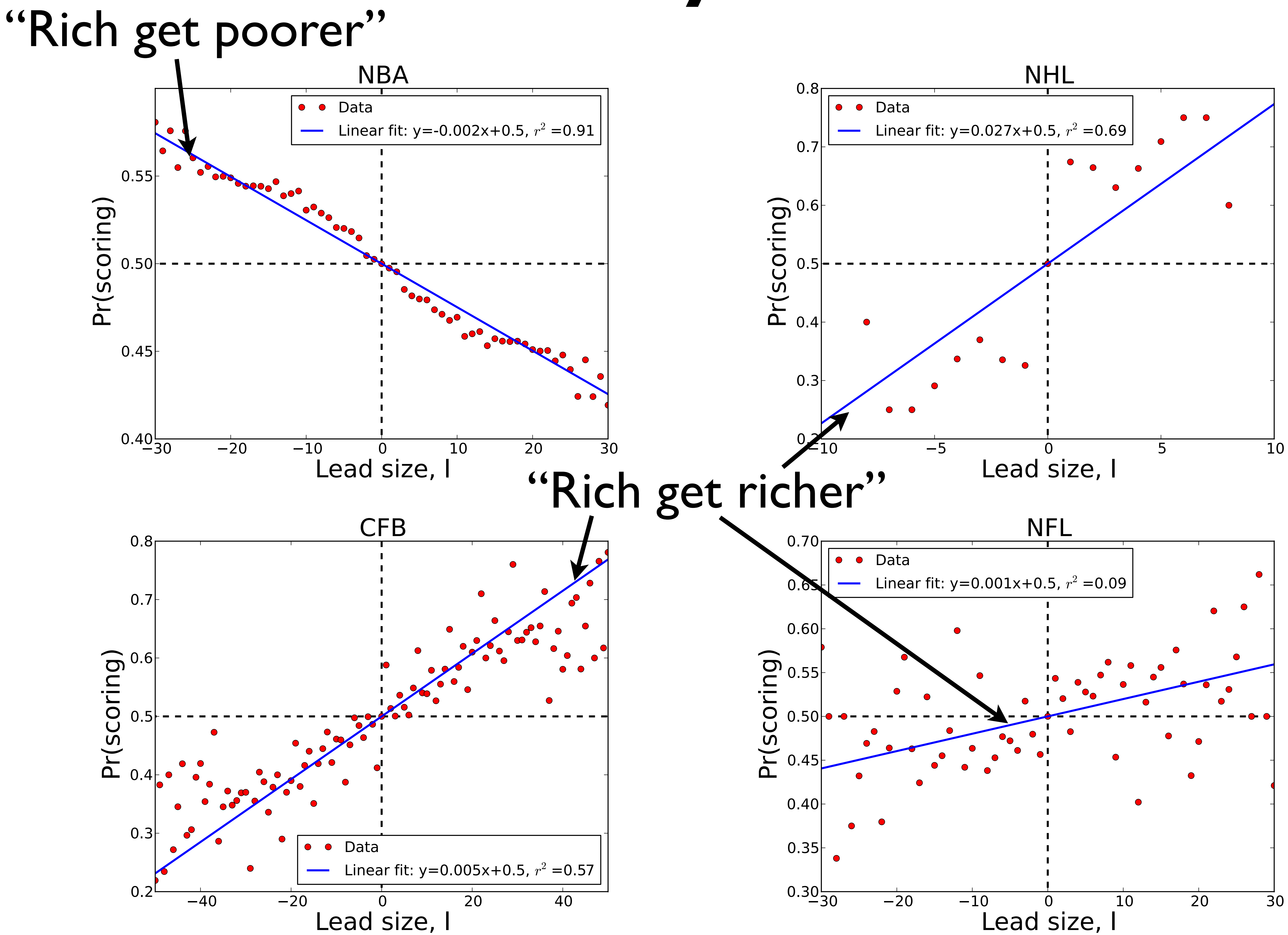
Maximize w.r.t. β

Balance



NBA $\hat{\beta} = 453.75$
NFL $\hat{\beta} = 1.86$
CFB $\hat{\beta} = 1.14$
NHL $\hat{\beta} = 0.63$
Density
0.0
0.2
0.4
0.6
0.8
1.0
Scoring bias, c

Lead dynamics



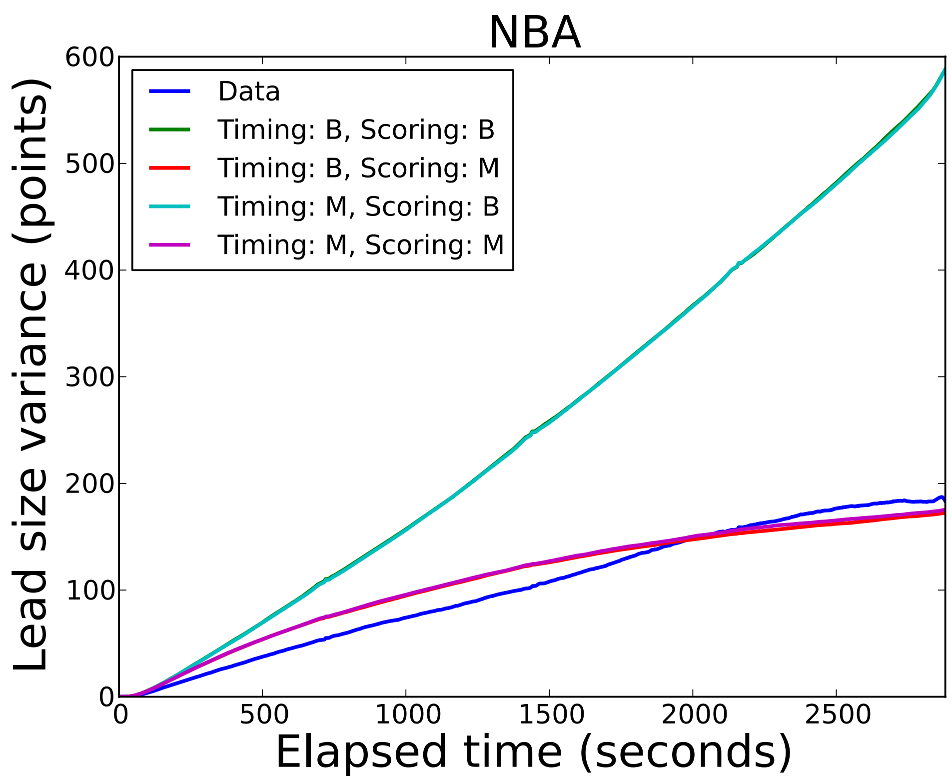
Points

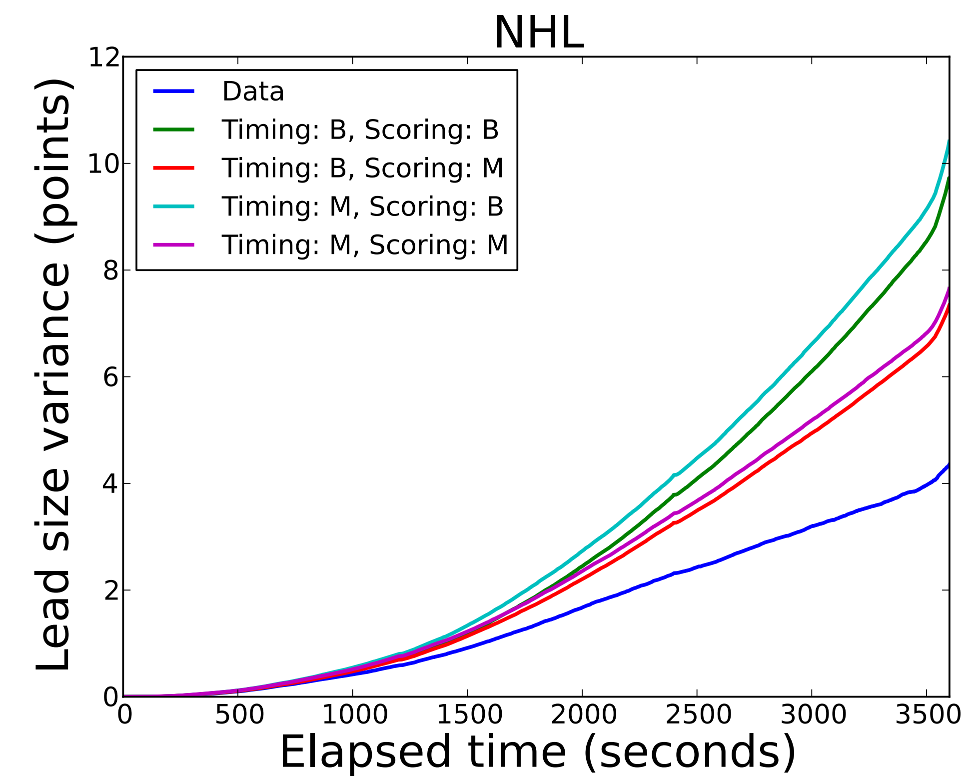
	fraction		
point value	NFL	CFB	NBA
1	-	-	0.097
2	0.009	0.014	0.738
3	0.290	0.017	0.161
6	0.032	0.071	-
7	0.427	0.514	-
8	0.016	0.018	-

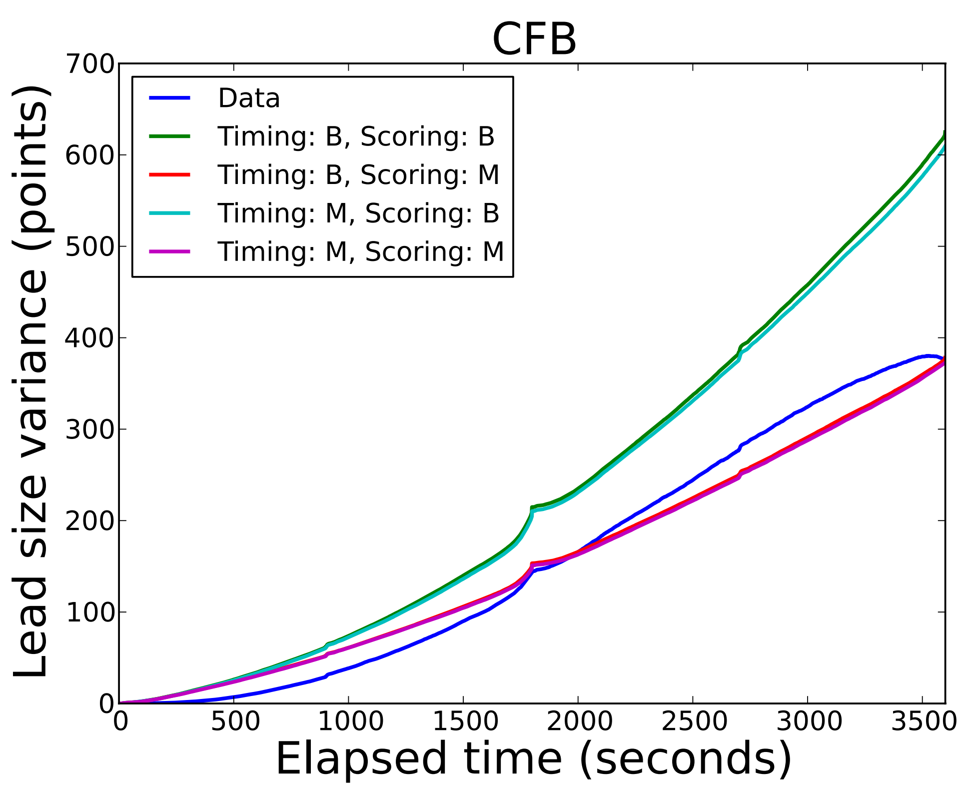
Non-parametric simulation

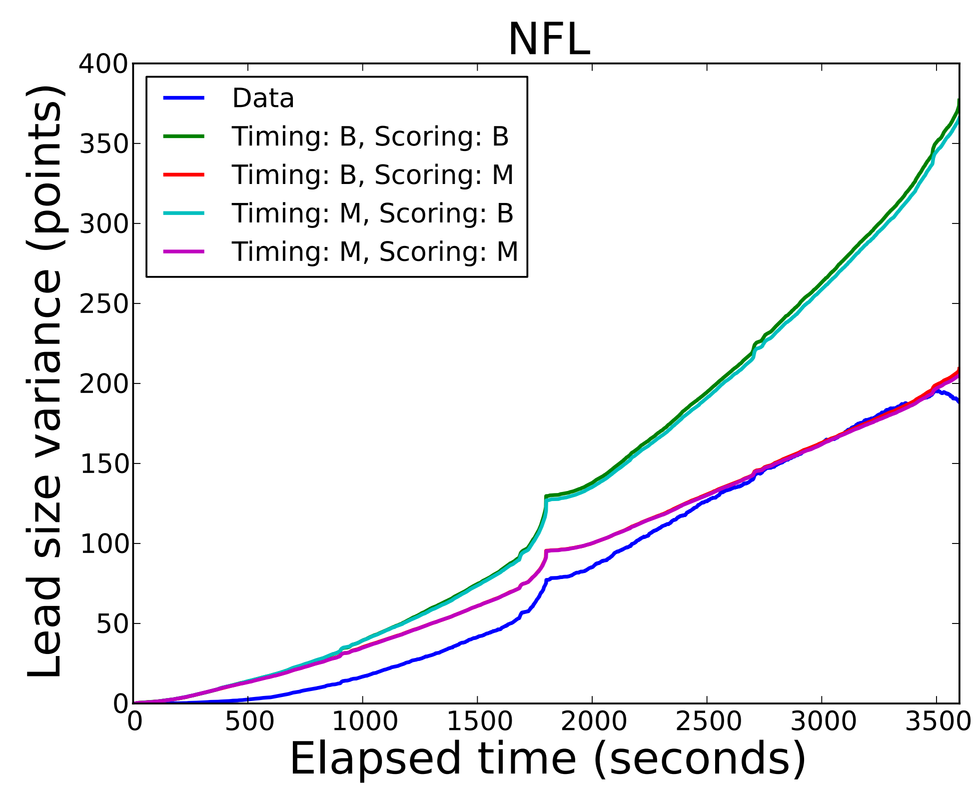
$t, S_r, S_b \leftarrow 0$
while $t \leq T$ **do**
  $t \leftarrow \text{t} + \text{get\_next\_time()}$
  $w \leftarrow \text{get\_winner()}$
  $p \leftarrow \text{get\_points()}$
  if $\text{w} = S_r$ **then**
    $S_r \leftarrow S_r + p$
  else
    $S_b \leftarrow S_b + p$
  end if
end while

Simulation

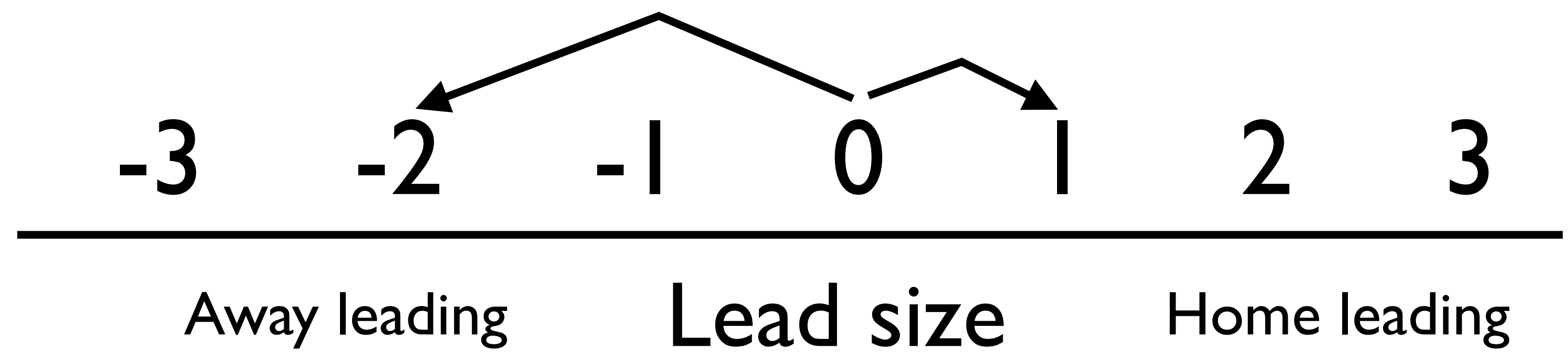

NBA
Data
Timing: B, Scoring: B
Timing: B, Scoring: M
Timing: M, Scoring: B
Timing: M, Scoring: M
Lead size variance (points)
0
100
200
300
400
500
600
0
500
1000
1500
2000
2500
Elapsed time (seconds)


NHL
Data
Timing: B, Scoring: B
Timing: B, Scoring: M
Timing: M, Scoring: B
Timing: M, Scoring: M
Lead size variance (points)
0
2
4
6
8
10
12
0
500
1000
1500
2000
2500
3000
3500
Elapsed time (seconds)


CFB
Data
Timing: B, Scoring: B
Timing: B, Scoring: M
Timing: M, Scoring: B
Timing: M, Scoring: M
Lead size variance (points)
0
100
200
300
400
500
600
700
0
500
1000
1500
2000
2500
3000
3500
Elapsed time (seconds)


NFL
Data
Timing: B, Scoring: B
Timing: B, Scoring: M
Timing: M, Scoring: B
Timing: M, Scoring: M
Lead size variance (points)
0
50
100
150
200
250
300
350
400
0
500
1000
1500
2000
2500
3000
3500
Elapsed time (seconds)

Outcome prediction



Markov chain state space

$$
P = \begin{array}{c|ccccc}
 & -2 & -1 & 0 & 1 & 2 \\
\hline
-2 & 0 & 0.3 & 0.4 & 0.3 & 0 \\
-1 & .. & 0 & & & \\
0 & & & 0 & & \\
1 & & & & 0 & \\
2 & & & & & 0
\end{array}
$$

Outcome prediction

$$
P = \begin{array}{c|ccccc}
 & \text{-2} & \text{-1} & 0 & 1 & 2 \\
\hline
\text{-2} & 0 & 0.3 & 0.4 & 0.3 & 0 \\
\text{-1} & .. & 0 & & & \\
0 & & & 0 & & \\
1 & & & & 0 & \\
2 & & & & & 0
\end{array}
$$

$$P_{ij} = \Pr(\text{scoring} \mid l)\,\Pr(\text{points } = p)$$

Probability of lead transitioning from i to j
at next scoring event

Outcome prediction

After each event:

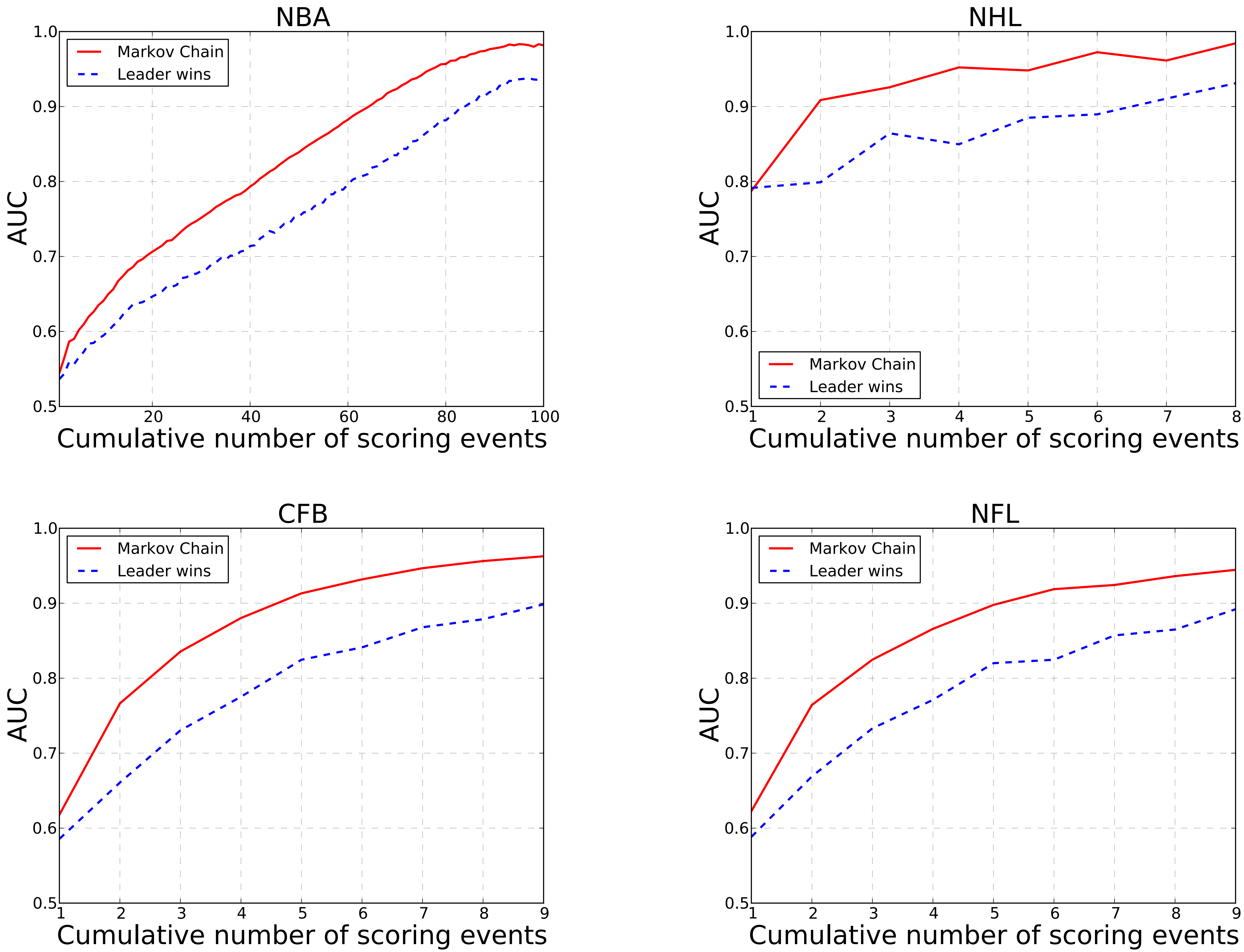
1. Estimate remaining number of scoring events

$$n = \sum_{i=t}^{T} \Pr(\text{event})(t)$$

2. Compute the probability lead ends in state > 0

$$\Pr(\text{team } r \text{ wins} \mid l, n) = \sum_{j=1}^{k} P_{lj}^{n}$$

Outcome prediction



Conclusions

- Global model of competition?
- Tempo follows a Poisson process
- First order Markov process captures nearly all scoring dynamics
- Competitions are predictable

The end

Thanks for listening