Football fever: goal distributions in football

Elmar Bittner
Wolfhard Janke
Andreas Nußbaumer
Universität Leipzig,
Leipzig, Germany

Martin Weigel
Heriot-Watt University,
Edinburgh, Scotland

What is the result of a football game?

1. Fußball-Bundesliga 2004/05
(first football league in Germany)

Away teams

	FC B	FC 0	SV W	Hert	VfB	Baye	Boru	Hamb	VfL	Hann	FSV	1.FC	Armi	1.FC	Boru	VfL	Hans	SC F
FC Bayern München		0:1	1:0	1:1	2:2	2:0	5:0	3:0	2:0	3:0	4:2	3:1	1:0	6:3	2:1	3:1	3:1	3:1
FC 04 Schalke	1:0		2:1	1:3	3:2	3:3	1:2	1:2	3:0	1:0	2:1	2:1	2:1	4:1	3:2	3:2	0:2	1:1
SV Werder Bremen	1:2	1:0		0:1	1:2	2:2	2:0	1:1	1:2	3:0	0:0	1:1	3:0	4:1	2:0	4:0	3:2	4:1
Hertha/BSC Berlin	0:0	4:1	1:1		0:0	3:1	0:1	4:1	3:1	0:0	1:1	1:1	3:0	2:1	6:0	2:2	1:1	3:1
VfB Stuttgart	1:3	3:0	1:2	1:0		3:0	2:0	2:0	0:0	1:0	4:2	1:1	2:1	2:4	1:0	5:2	4:0	1:0
Bayer 04 Leverkusen	4:1	0:3	2:1	3:3	1:1		0:1	3:0	2:1	2:1	2:0	2:0	3:2	2:2	5:1	4:0	3:0	4:1
Borussia Dortmund	2:2	0:1	1:0	2:1	0:2	1:0		0:2	1:2	1:1	3:0	4:2	1:1	2:2	1:1	1:0	2:1	2:0
Hamburger SV	0:2	1:2	1:2	2:1	2:1	1:0	2:3		3:1	0:2	2:1	2:1	0:2	4:3	0:0	0:1	3:0	4:0
VfL Wolfsburg	0:3	3:0	2:3	2:3	3:0	2:2	1:2	1:0		1:0	4:3	2:1	5:0	0:1	2:1	3:0	4:0	0:1
Hannover 96	0:1	1:0	1:4	0:1	0:0	0:3	1:3	2:1	3:0		2:0	3:1	0:1	1:0	2:1	3:0	0:1	2:2
FSV Mainz 05	2:4	2:1	2:1	0:3	2:3	2:0	1:1	2:1	0:2	2:0		3:2	0:0	0:1	1:1	1:0	3:1	5:0
1.FC Kaiserslautern	0:4	2:0	1:2	0:2	2:3	0:0	1:0	2:1	0:0	0:2	2:0		2:1	1:3	1:0	1:2	2:1	3:0
Arminia Bielefeld	3:1	0:2	2:1	1:0	0:2	1:0	1:0	3:4	1:2	0:1	1:1	0:2		3:1	0:0	1:2	1:1	3:1
1.FC Nürnberg	2:2	0:2	1:2	0:0	1:1	2:4	2:2	1:3	4:0	1:1	1:2	1:3	1:2		0:0	2:1	3:0	3:0
Borussia Mönchengladbach	2:0	1:3	3:1	0:0	2:0	1:1	2:3	1:3	1:0	0:2	1:1	2:0	1:0	2:1		2:2	2:2	3:2
VfL 1848 Bochum	1:3	0:2	1:4	2:2	2:0	2:2	2:2	1:2	5:1	1:0	2:6	1:1	1:1	3:1	3:0		0:1	3:1
Hansa Rostock	0:2	2:2	0:4	2:1	2:1	0:2	1:1	0:6	1:2	1:3	2:0	2:3	1:1	0:2	0:0	3:1		0:0
SC Freiburg	0:1	2:3	0:6	1:3	2:0	1:3	2:2	1:1	1:0	0:0	1:2	1:2	2:3	2:3	1:1	1:1	0:0	

18 teams, i.e. $18 \cdot 17 = 306$ games

Home teams

What results do we have and where do they come from?

European leagues:

- Belgium
- Bulgaria
- GDR
- Germany 1st/2nd
- England
- France
- Italy
- Netherlands
- Austria
- Portugal
- Rumania
- Russia
- Scotland
- Spain
- Czechoslovakia
- Hungary

Cup matches:

- Germany
- UEFA

Women leagues:

- Austria
- Germany

Tournaments:

- European Championships
- World Championchips

Data came from the Internet:

- **http://www.fussballdaten.de**
- **http://www.fussballportal.de**
- **http://www.nordostfussball.de**
- **http://www.sportergebnise.de**

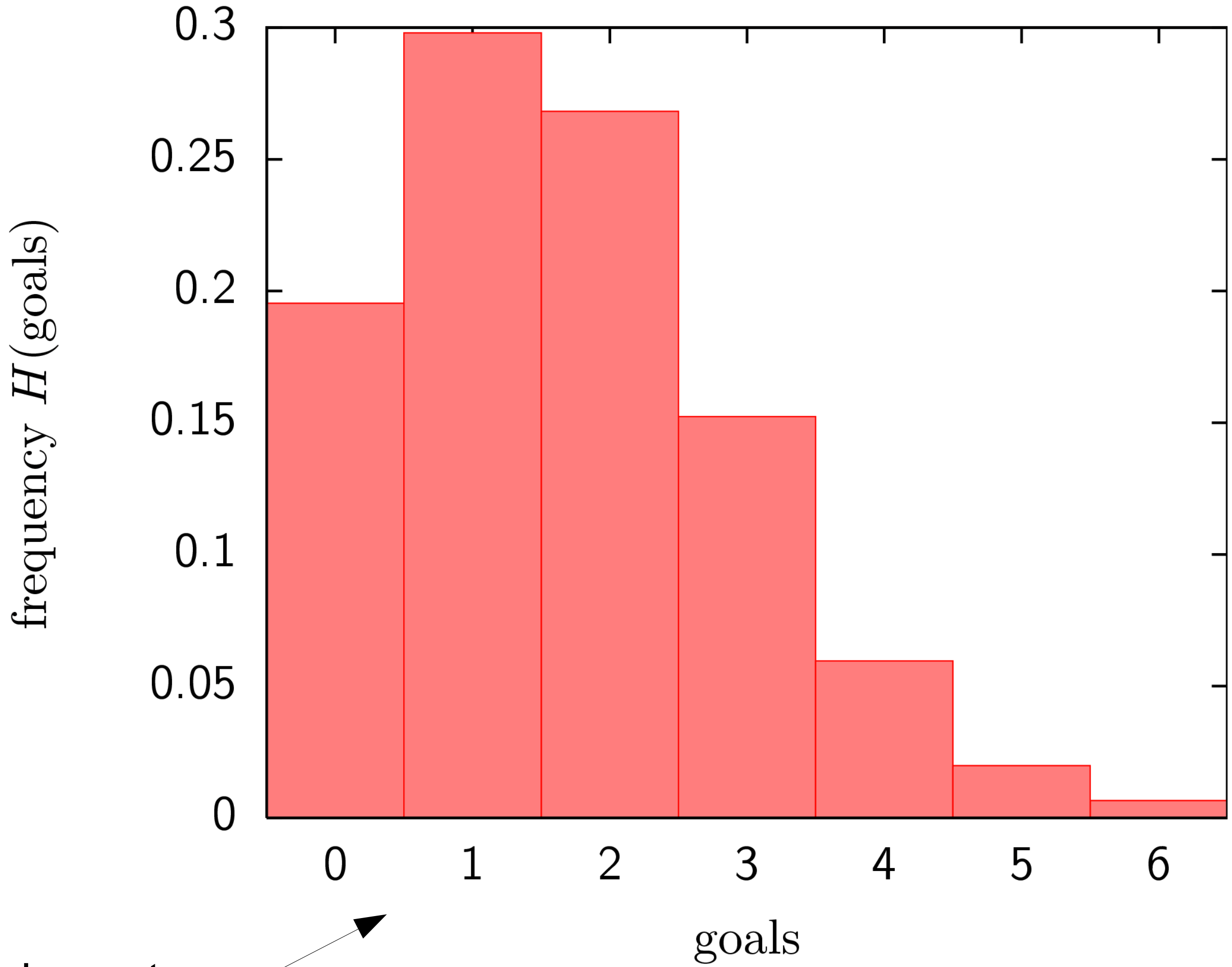
perl, awk, grep, ...

Simple list: <home> <away>

1 2
2 2
0 2
0 3
1 1
4 1
...

A “typical” result:

1. Fußball-Bundesliga
2004/05 (Germany)



home team

Goals	Matches
0	60
1	90
2	82
3	48
4	18
5	6
6	2

Question: what kind of distribution is this?

Simplest possible idea:

Goals are scored with a constant probability (in time)!

The scores of the home and the away team don't influence each other!

N number of time steps (?)

p probability to score a goal during one time step

goals

Binomial distribution

$$P_{N,p}(k) = \binom{N}{k} p^k (1-p)^{N-k}$$

$$N p = \lambda \rightarrow p = \frac{\lambda}{N}$$

Poisson distribution

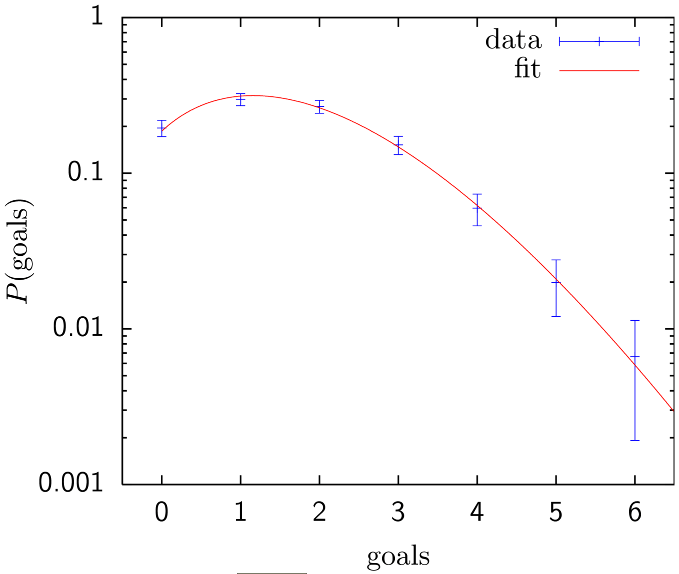
$$\lim_{N\rightarrow\infty} P_{N,p}(k) = P_\lambda(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

average number of goals: $\langle P_\lambda(k) \rangle_k = \lambda$

Does the Poisson distribution work?

(At least it worked with horses and Preussian soldiers!)

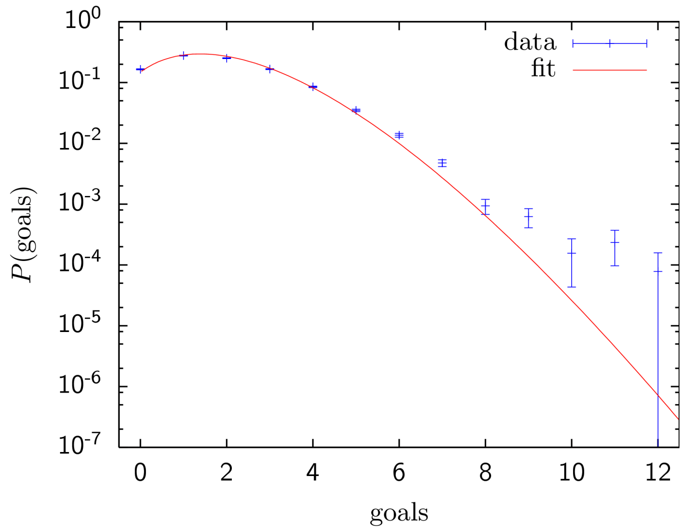
1. Fußball-Bundesliga 2004/05 (Germany)



$\chi^2/\text{d.o.f} = 0.1$

306 games

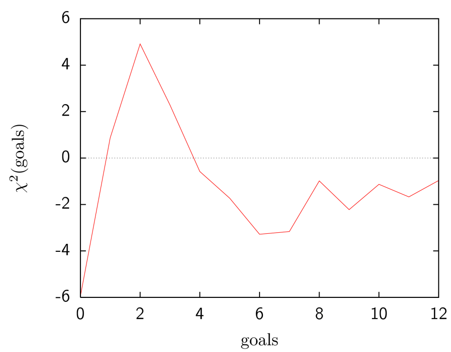
1. Fußball-Bundesliga1963-2004 (Germany)



$\chi^2/\text{d.o.f} = 8.2$

12794 games

Errors from bootstrap are wrong!
More data is needed!



the fit is bad for small numbers of goals

Conclusion: it may be good for a single (?) game but otherwise pretty useless

C'mon!

Maybe a model can help?

Ingredients:

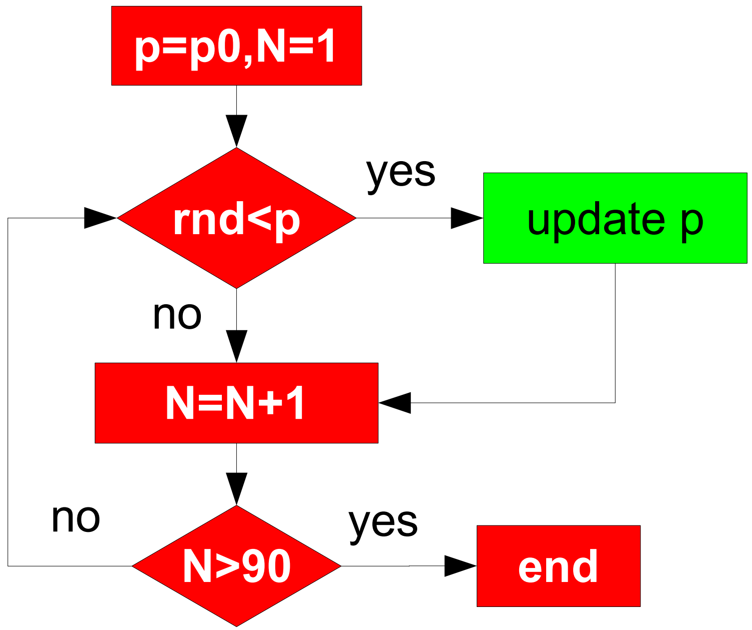
1. Starting probability
2. Probability to score a goal depends on the actual score

3 different updates if a goal is scored:

	home scores	away scores
model “A”:	$p_h \leftarrow p_h + \kappa_h$	$p_a \leftarrow p_a + \kappa_a$
model “B”:	$p_h \leftarrow p_h \cdot \kappa_h$	$p_a \leftarrow p_a \cdot \kappa_a$
model “C”:	$p_h \leftarrow p_h \cdot \kappa_h$ $p_a \leftarrow p_a / \kappa_a$	$p_h \leftarrow p_h / \kappa_h$ $p_a \leftarrow p_a \cdot \kappa_a$

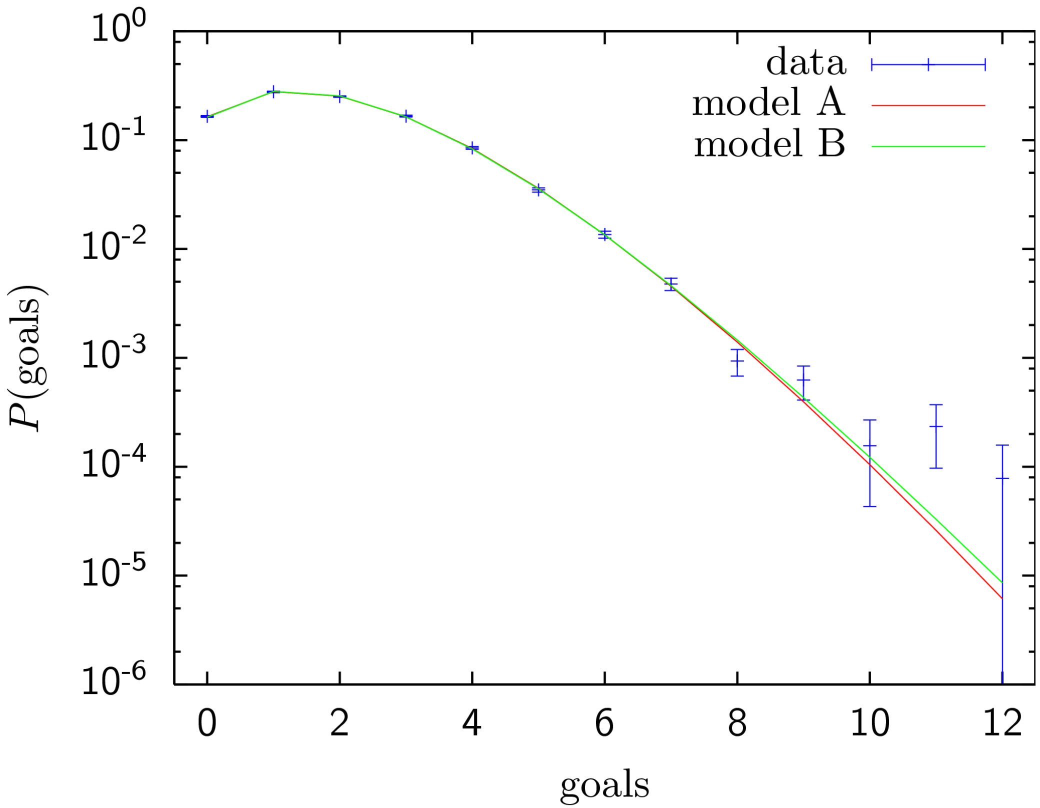
κ : additional motivation
“success succeeds”

Algorithm:



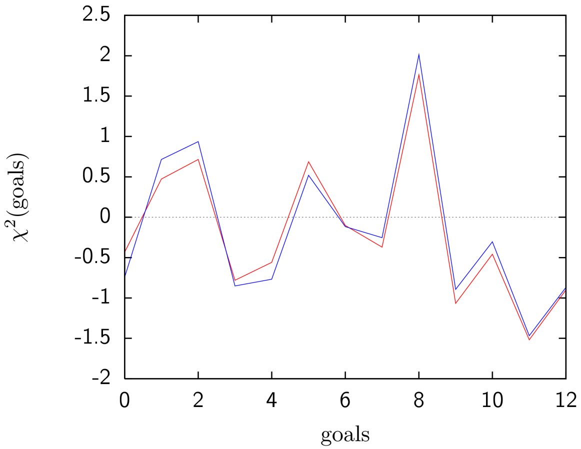
Fitting of the model

1. Fußball-Bundesliga1963-2004 (Germany)



Model “A”: $\chi^2/\text{d.o.f} = 0.91$

Model “B”: $\chi^2/\text{d.o.f} = 1.04$



We don't even need to simulate the model!

$$P_N(n) = [1-p(n)]\,P_{N-1}(n) + p(n-1)\,P_{N-1}(n-1)$$

simple recursion

no goal at time n — goal at time n

A: $p(n) = p_0 + \kappa\, n$

B: $p(n) = p_0 \kappa^n$

...

technical: fill matrix from bottom to top in O(n N) steps

What distribution is described by model “A”, “model B”, ...

take recursion relation → transform in difference equation

→ perform continuum limit $N \rightarrow t$

→ transform in differential equation

→ solve differential equation using $p(n)$

(first three steps:) independent of a specific $p(n)$

Finally,
for **model “A”**

$$P_{p_0,\kappa,t}(n) = e^{-p_0 t} \frac{\Gamma(p_0/\kappa + n)}{n!\,\Gamma(p_0/\kappa)} \left(1 - e^{-\kappa t}\right)^n$$

compare

negative binomial distribution

$$P_{r,p}(n) = \frac{\Gamma(r+n)}{n!\,\Gamma(r)} p^n (1-p)^r$$

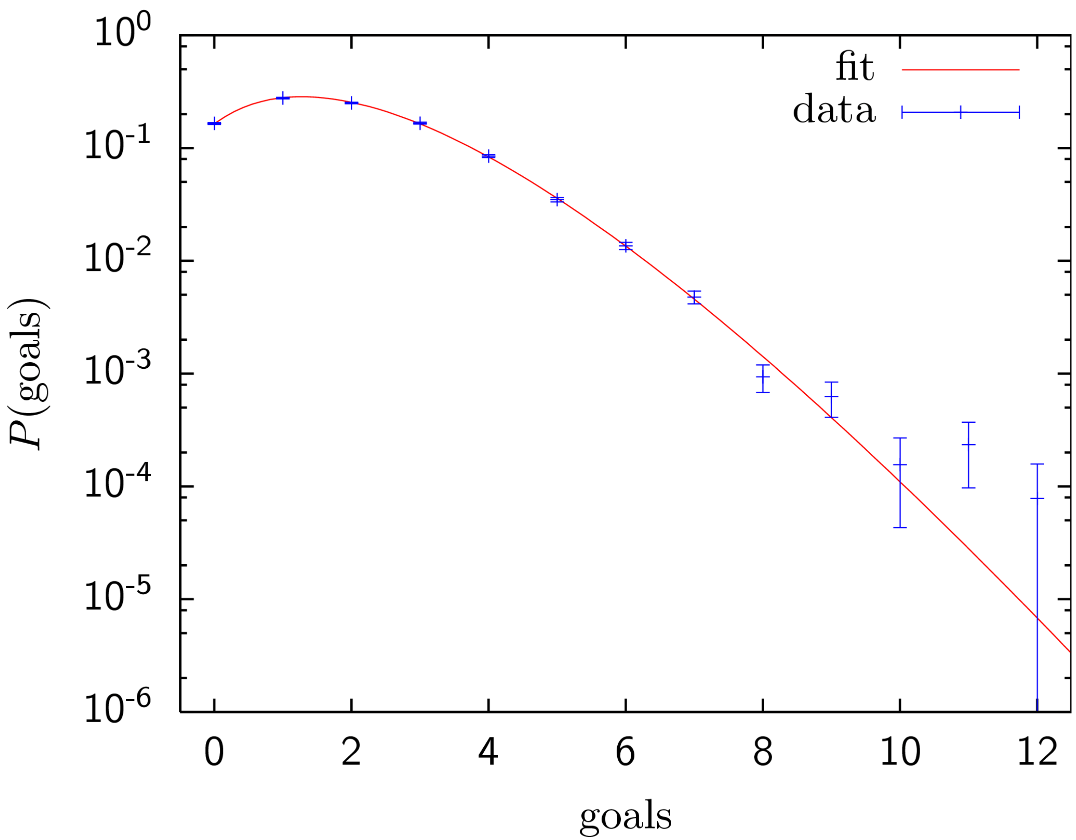
$$p = 1 - e^{-\kappa t} \qquad r = \frac{p_0}{\kappa}$$

No closed expressions for model ”B” and model “C” so far!

The **negative binomial distribution** describes a **process** with:

- n+r Bernoulli trials (0,1)
- success probability p
- n times success
- r times failure
- last attempt is a failure

Fitting of the **n**egative **b**inomial **d**istribution (NBD)



NBD: $\chi^2/\text{d.o.f} = 0.95$

Are we the first that tried to fit a NBD to football results?

Not exactly! See **M.J. Moroney: "Facts from Figures" (1956)**

What else did we fit? General extreme value (GEV) distributions!

$$P_{\xi,\mu,\sigma}(n) = \frac{1}{\sigma}\left(1+\xi\frac{n-\mu}{\sigma}\right)^{1-1/\xi}\exp\left[-\left(1+\xi\frac{n-\mu}{\sigma}\right)^{-1/\xi}\right] \qquad \xi \neq 0$$

$$P_{\mu,\sigma}(n) = \frac{1}{\sigma}\exp\left[-\exp\left(-\frac{n-\mu}{\sigma}\right)-\frac{n-\mu}{\sigma}\right] \qquad \xi \to 0$$

shape parameter ξ
scale parameter σ
location parameter μ

most important, controls the overall shape of the function:

- $\xi < 0$: Weibull
- $\xi > 0$: Fréchet
- $\xi = 0$: Gumbel

Why did we fit GEV? We saw it in Greenhough et al.: "Football goal distributions and extremal statistics", Physica A **316** (2002) 615-624

But 2006: Bertin and Clusel: "Generalised extreme value statistics and sum of correlated variables", cond-mat/0601189

⇨ miscroscopical model with **feedback** can lead to GEV

We did not only fit the score of the home team but also:

- the score of the away team
- the sum of the scores → $P^{\Sigma}(K)=\sum_{k=0}^{K} P^{h}(k)\,P^{a}(K-k)$
- and the difference of the scores → $P^{\Delta}(d)=\sum_{k=0}^{\infty} P^{h}(k+d)\,P^{a}(k)$

e.g.: $h=2, a=1$ then $K=h+a=2+1=3$

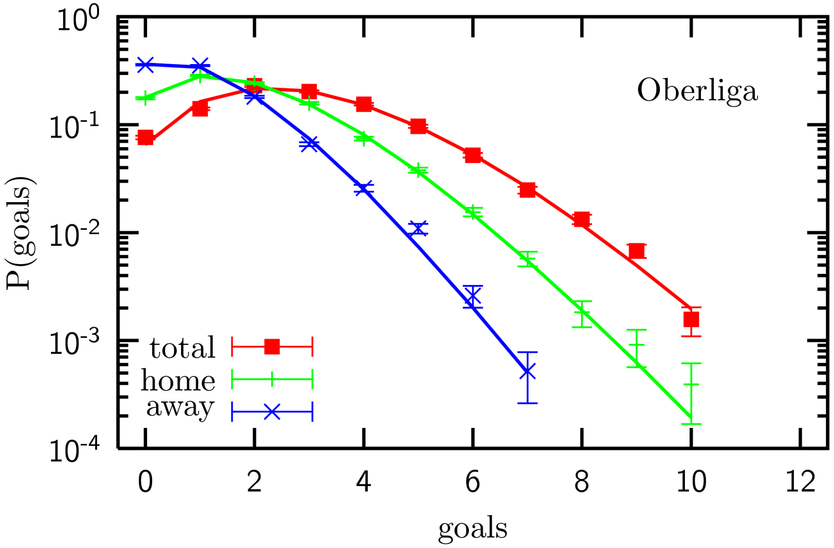
and $d=h-a=2-1=1$

To summarize:

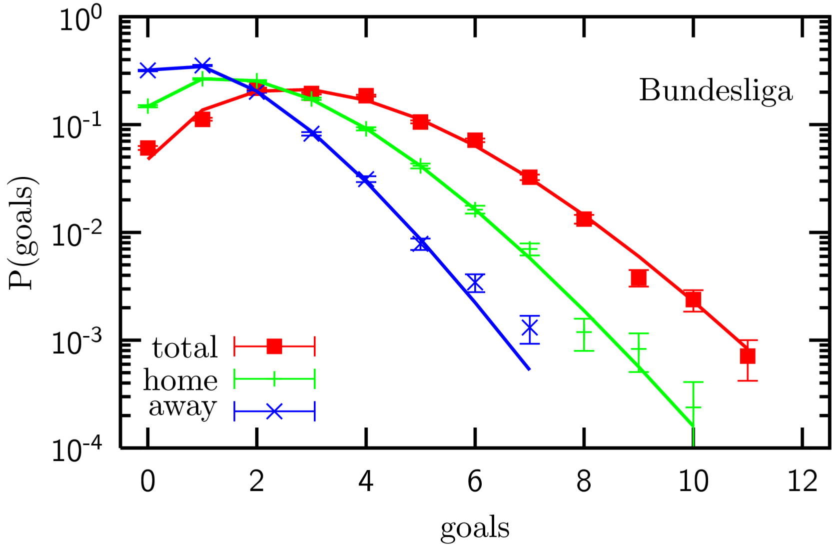
	home/away	sum	difference
Poisson:	$P_{\lambda}(k)=\frac{\lambda^{k}}{k!}e^{-\lambda}$	$P^{\Sigma}_{\lambda_h,\lambda_a}(K)=\frac{(\lambda_h+\lambda_a)^{K}}{K!}e^{-(\lambda_h+\lambda_a)}$	$P^{\Delta}_{\lambda_h,\lambda_a}(d)=\left(\frac{\lambda_h}{\lambda_a}\right)^{d/2} I_d(2\sqrt{\lambda_h\lambda_a})\; e^{-(\lambda_h+\lambda_a)}$
NBD:	$P_{r,p}(n)=\frac{\Gamma(r+n)}{n!\,\Gamma(r)}p^{n}(1-p)^{r}$	$P^{\Sigma}_{p_h,r_h,p_a,r_a}(K)=(1-p_h)^{r_h}(1-p_a)^{r_a}p_a^{K}\frac{\Gamma(r_a+K)}{K!\,\Gamma(r_a)}F\left(-K,r_h;1-K-r_a;\frac{p_h}{p_a}\right)$	$P^{\Sigma}_{p_h,r_h,p_a,r_a}(d)=(1-p_h)^{r_h}(1-p_a)^{r_a}p_h^{d}\frac{\Gamma(r_h+d)}{d!\,\Gamma(r_h)}F(r_h+d,r_a;1+d;p_h p_a)$
Weibull/Frechet:	$P_{\xi,\mu,\sigma}(n)=\frac{1}{\sigma}\left(1+\xi\frac{n-\mu}{\sigma}\right)^{1-1/\xi}\exp\left[-\left(1+\xi\frac{n-\mu}{\sigma}\right)^{-1/\xi}\right]$	no closed expressions for sum/difference	
Gumbel:	$P_{\mu,\sigma}(n)=\frac{1}{\sigma}\exp\left[-\exp\left(-\frac{n-\mu}{\sigma}\right)-\frac{n-\mu}{\sigma}\right]$		
Model “A”/”B”:	approximation for finite N can be found		

Comparison GDR vs. West Germany

Oberliga 49/50-90/91 (~7.700 matches)



Bundesliga 63/64-90/91 (~ 8.400 matches)



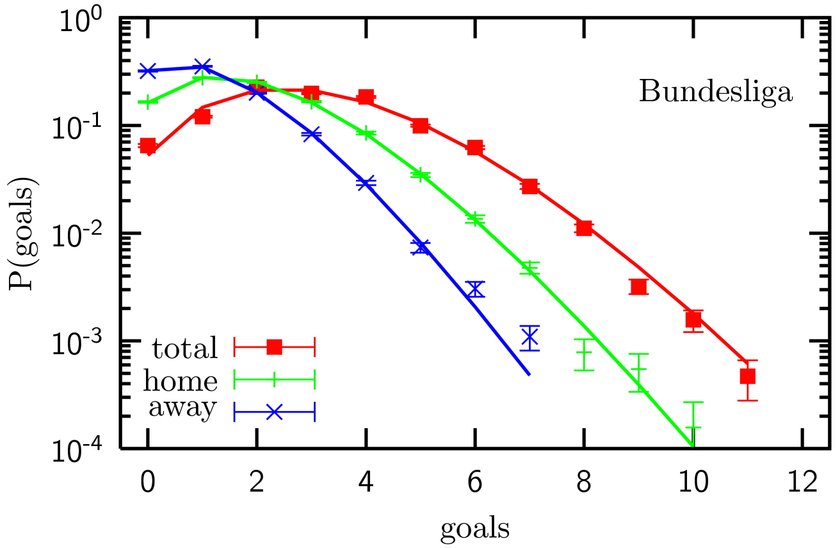
		Oberliga		Bundesliga	
		Home	Away	Home	Away
Poisson	λ	1.85 ± 0.02	1.05 ± 0.01	2.01 ± 0.02	1.17 ± 0.01
	χ^2/d.o.f.	12.5	12.8	6.53	7.31
NBD	p	0.17 ± 0.01	0.14 ± 0.01	0.11 ± 0.01	0.10 ± 0.01
	r	9.06 ± 0.88	6.90 ± 0.84	15.9 ± 2.10	11.3 ± 1.84
	p_0	0.0191	0.0112	0.0213	0.0126
	κ	0.0021	0.0016	0.0013	0.0011
	χ^2/d.o.f.	0.99	4.09	0.68	2.29
GEV	ξ	-0.05 ± 0.01	0.02 ± 0.01	-0.09 ± 0.01	-0.01 ± 0.01
	μ	1.12 ± 0.02	0.49 ± 0.02	1.28 ± 0.02	0.58 ± 0.02
	σ	1.30 ± 0.02	0.90 ± 0.02	1.36 ± 0.02	0.96 ± 0.02
	χ^2/d.o.f.	1.93	5.04	1.83	4.74
Gumbel	μ	1.12 ± 0.02	0.48 ± 0.02	1.28 ± 0.02	0.59 ± 0.01
	σ	1.25 ± 0.01	0.92 ± 0.01	1.25 ± 0.01	0.95 ± 0.01
	χ^2/d.o.f.	4.13	4.65	12.9	4.06

all fits (GDR/West Germany) are NBD

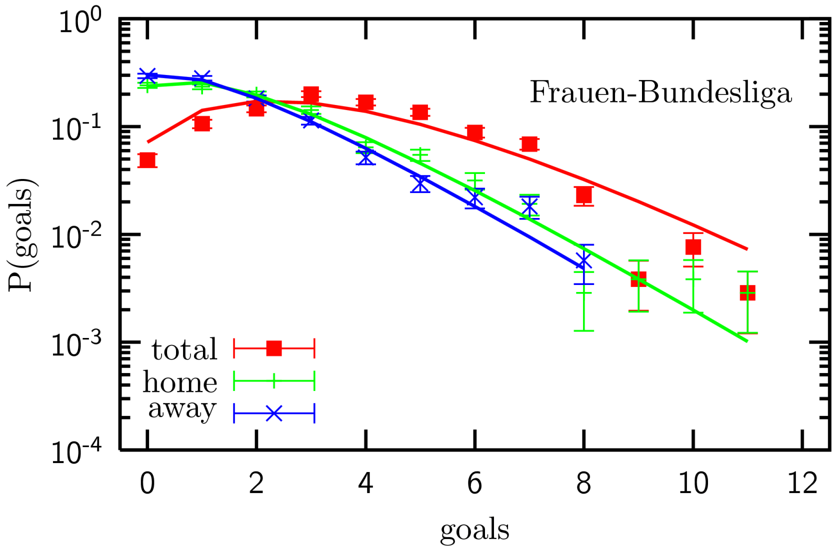
Crux: GDR goals are more encouraging

Bundesliga: men vs. women

Bundesliga men 63/64-04/05 (12800 matches)



Bundesliga women 97/98-04/05 (1050 matches)

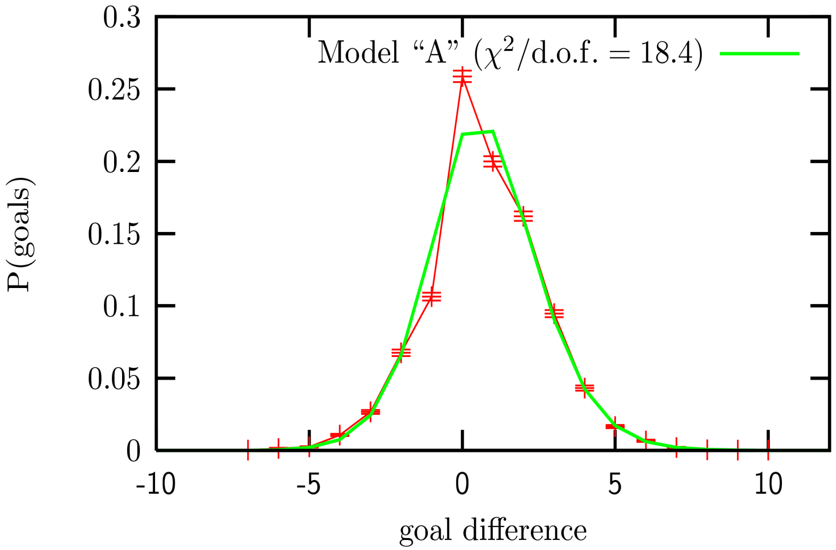


		Bundesliga		Frauen-Bundesliga	
		Home	Away	Home	Away
Poisson	λ	1.91 ± 0.01	1.16 ± 0.01	1.78 ± 0.04	1.36 ± 0.04
	χ^2/d.o.f.	9.21	9.13	14.6	14.4
NBD	p	0.11 ± 0.01	0.09 ± 0.01	0.45 ± 0.03	0.46 ± 0.03
	r	16.24 ± 1.82	12.08 ± 1.69	2.38 ± 0.24	1.97 ± 0.22
	p_0	0.0202	0.0125	0.0160	0.0133
	κ	0.0012	0.0010	0.0067	0.0068
	χ^2/d.o.f.	1.08	2.22	2.32	1.37
GEV	ξ	-0.10 ± 0.01	-0.02 ± 0.01	0.04 ± 0.04	0.25 ± 0.07
	μ	1.17 ± 0.02	0.57 ± 0.01	0.83 ± 0.08	0.77 ± 0.07
	σ	1.33 ± 0.01	0.96 ± 0.01	1.49 ± 0.06	1.18 ± 0.05
	χ^2/d.o.f.	3.43	7.95	3.40	1.55
Gumbel	μ	1.18 ± 0.01	0.58 ± 0.01	0.81 ± 0.08	0.58 ± 0.07
	σ	1.21 ± 0.01	0.94 ± 0.01	1.53 ± 0.05	1.31 ± 0.05
	χ^2/d.o.f.	24.5	7.26	3.17	4.09

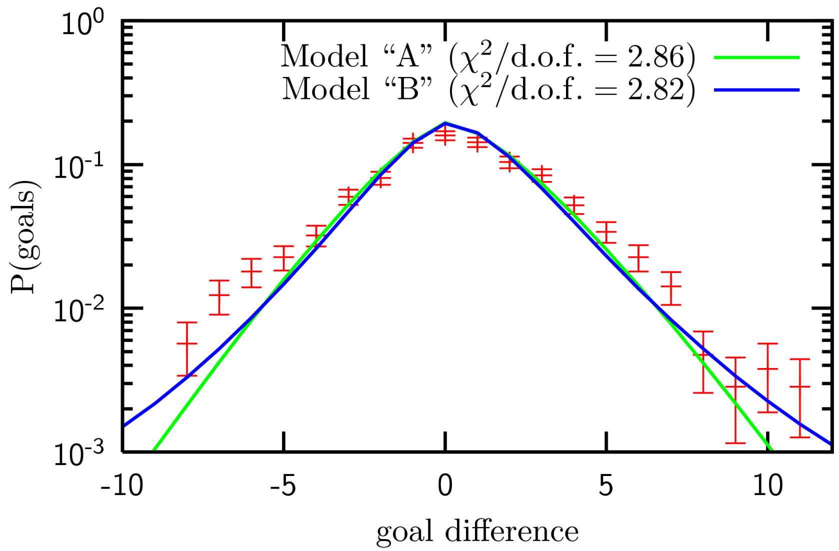
all fits (men/women) are NBD

Sum and difference:

Bundesliga men 63/64-04/05 (12800 matches)



Bundesliga women 97/98-04/05 (1050 matches)



		Bundesliga 04/05	Bundesliga 90/91	Oberliga	Women
Poisson	λ_h	1.91 ± 0.01	2.01 ± 0.02	1.85 ± 0.02	1.78 ± 0.04
	λ_a	1.16 ± 0.01	1.17 ± 0.01	1.05 ± 0.01	1.36 ± 0.04
Home	χ^2_h/d.o.f.	9.21	6.53	12.5	14.6
Away	χ^2_a/d.o.f.	9.13	7.31	12.8	14.4
Total	χ^2_Σ/d.o.f.	10.7	15.9	16.3	10.4
Difference	χ^2_Δ/d.o.f.	67.6	578	474	20.2
NBD	p_h	0.11 ± 0.01	0.11 ± 0.01	0.17 ± 0.01	0.45 ± 0.03
	r_h	16.24 ± 1.82	15.9 ± 2.10	9.06 ± 0.82	2.38 ± 0.24
	p_a	0.09 ± 0.01	0.10 ± 0.01	0.14 ± 0.01	0.46 ± 0.03
	r_h	12.08 ± 1.69	11.3 ± 1.84	6.90 ± 0.84	1.97 ± 0.22
Home	χ^2_h/d.o.f.	1.08	0.68	0.99	2.32
Away	χ^2_a/d.o.f.	2.22	2.29	4.09	1.37
Total	χ^2_Σ/d.o.f.	25.1	17.3	8.31	18.9
Difference	χ^2_Δ/d.o.f.	23.9	18.0	7.16	3.55

		Bundesliga 04/05	Bundesliga 90/91	Oberliga	Women
Model "A"	$p_{0,h}$	0.0199 ± 0.0002	0.0210 ± 0.0002	0.0188 ± 0.0002	0.0159 ± 0.0005
	κ_h	0.0015 ± 0.0001	0.0016 ± 0.0002	0.0024 ± 0.0002	0.0070 ± 0.0005
	$p_{0,a}$	0.0125 ± 0.0002	0.0125 ± 0.0001	0.0112 ± 0.0001	0.0132 ± 0.0004
	κ_a	0.0012 ± 0.0001	0.0013 ± 0.0002	0.0018 ± 0.0002	0.0071 ± 0.0007
Home	χ^2_h/d.o.f.	1.01	0.68	1.07	2.28
Away	χ^2_a/d.o.f.	2.31	2.37	4.23	1.44
Total	χ^2_Σ/d.o.f.	16.6	11.5	5.33	12.4
Difference	χ^2_Δ/d.o.f.	18.6	14.0	5.63	2.86
Model "B"	$p_{0,h}$	0.0200 ± 0.0002	0.0211 ± 0.0002	0.0189 ± 0.0002	0.0166 ± 0.0005
	κ_h	1.0679 ± 0.0060	1.0695 ± 0.0072	1.1115 ± 0.0083	1.3146 ± 0.0303
	$p_{0,a}$	0.0125 ± 0.0001	0.0125 ± 0.0002	0.0112 ± 0.0001	0.0138 ± 0.0004
	κ_a	1.0932 ± 0.0106	1.1015 ± 0.0124	1.1526 ± 0.0149	1.4115 ± 0.0543
Home	χ^2_h/d.o.f.	1.25	0.71	0.75	3.24
Away	χ^2_a/d.o.f.	1.96	2.02	3.35	0.95
Total	χ^2_Σ/d.o.f.	16.9	11.8	5.40	13.5
Difference	χ^2_Δ/d.o.f.	18.4	13.8	5.26	2.82

sum, difference as consistency check

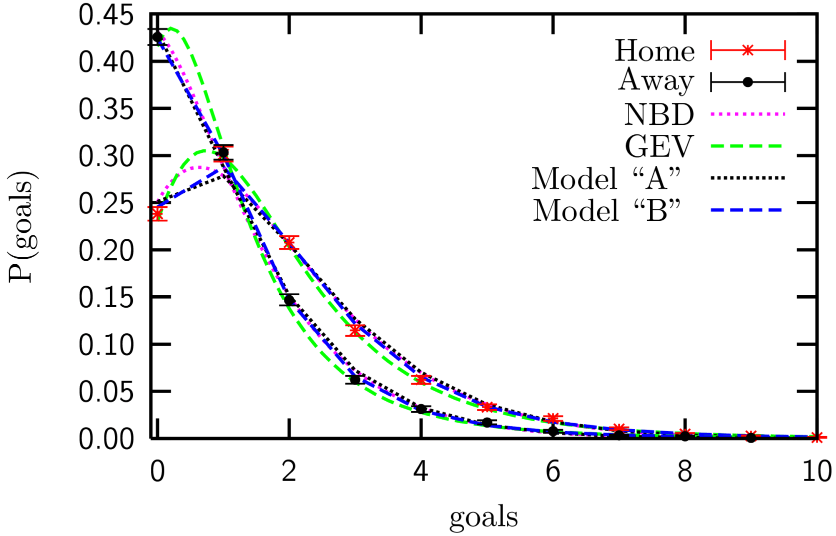
$$R = \frac{\mathrm{Cov}(k_h, k_a)}{\sigma_h \sigma_a}$$

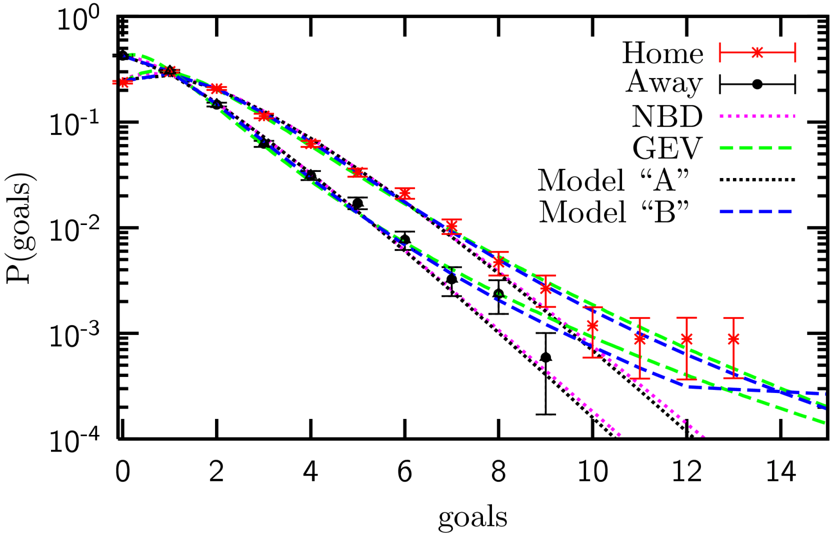
$$R_{\mathrm{BL}} = -0.031$$

$$R_{\mathrm{OL}} = -0.015$$

FIFA World Championships (TM)

Qualification phase (~ 3400 matches)





		Home	Away
Poisson	λ	1.53 ± 0.02	0.89 ± 0.01
	χ^2/d.o.f.	18.6	25.0
NBD	p	0.37 ± 0.02	0.38 ± 0.02
	r	3.04 ± 0.21	1.76 ± 0.12
	p_0	0.0154	0.0094
	κ	0.0051	0.0053
	χ^2/d.o.f.	2.67	2.02
GEV	ξ	0.11 ± 0.02	0.19 ± 0.02
	μ	0.86 ± 0.03	0.36 ± 0.03
	σ	1.21 ± 0.03	0.86 ± 0.02
	χ^2/d.o.f.	0.85	1.89
Gumbel	μ	0.80 ± 0.03	0.25 ± 0.03
	σ	1.31 ± 0.02	0.94 ± 0.02
	χ^2/d.o.f.	3.29	12.9
Model "A"	p_0	0.0152 ± 0.0003	0.0093 ± 0.0002
	κ	0.0053 ± 0.0003	0.0055 ± 0.0003
	χ^2/d.o.f.	2.88	2.19
Model "B"	p_0	0.0155 ± 0.0002	0.0095 ± 0.0002
	κ	1.2780 ± 0.0130	1.4775 ± 0.0343
	χ^2/d.o.f.	0.92	0.80

Summary by us:

Technical:
- Poisson fits bad (in all cases)
- NBD fits quite well (in the majority of cases)
- Heavy tails fit even better with GEV
- Model “A” can be compared to NBD
- Model “B” can fit large tails (as GEV)

Interpretation:
- Strongly different team strengths lead to fat tails
- Fat tails lead to large κ
- Oberliga vs. Bundesliga can be compared to women to men
- World Cup qualification can be compared to the women's league

Summary by Sepp Herberger

- “Das nächste Spiel ist immer das schwerste Spiel!”
- “Das Runde muss in das Eckige.”
- “Der Ball ist rund und das Spiel dauert 90 Minuten.”
- “Der nächste Gegner ist immer der schwerste.”
- “Der schnellste Spieler ist der Ball.”

More information:

xxx.lanl.gov

physics/0606016


lanl.arXiv.org > physics > physics/0606016

Search for (Help | Advanced search)

All papers Go!

Physics, abstract
physics/0606016

From: Elmar Bittner [view email]
Date: Thu, 1 Jun 2006 16:42:24 GMT (149kb)

Football fever: goal distributions and non-Gaussian statistics

Authors: Elmar Bittner, Andreas Nussbaumer, Wolfhard Janke, Martin Weigel
Comments: 24 pages, 7 figures
Subj-class: Physics and Society

Analyzing football score data with statistical techniques, we investigate how the not purely random, but highly co-operative nature of the game is reflected in averaged properties such as the probability distributions of scored goals for the home and away teams. As it turns out, especially the tails of the distributions are not well described by the Poissonian or binomial model resulting from the assumption of uncorrelated random events. Instead, a good effective description of the data is provided by less basic distributions such as the negative binomial one or the probability densities of extreme value statistics. To understand this behavior from a microscopical point of view, however, no waiting time problem or extremal process need be invoked. Instead, modifying the Bernoulli random process underlying the Poissonian model to include a simple component of self-affirmation seems to describe the data surprisingly well and allows to understand the observed deviation from Gaussian statistics. The phenomenological distributions used before can be understood as special cases within this framework. We analyzed historical football score data from many leagues in Europe as well as from international tournaments, including data from all past tournaments of the ``FIFA World Cup'' series, and found the proposed models to be applicable rather universally. In particular, here we analyse the results of the German women's premier football league and consider the two separate German men's premier leagues in the East and West during the cold war times and the unified league after 1990 to see how scoring in football and the component of self-affirmation depend on cultural and political circumstances.

Full-text: PostScript, PDF, or Other formats

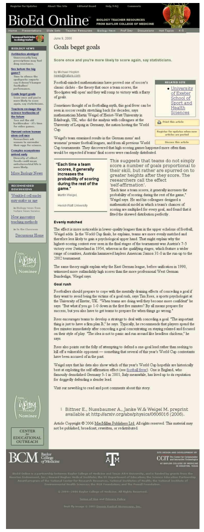
References and citations for this submission:
CiteBase (autonomous citation navigation and analysis)

Which authors of this paper are endorsers?

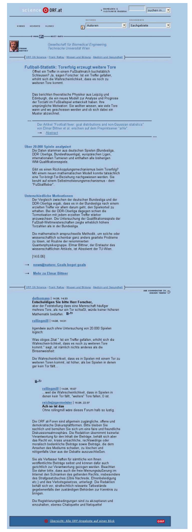
Links to: arXiv, form interface, /find, physics, /0606, /abs, previous, next, help (Access key information)

.DE addresses: Nearest e-print mirror site is http://de.arXiv.org/

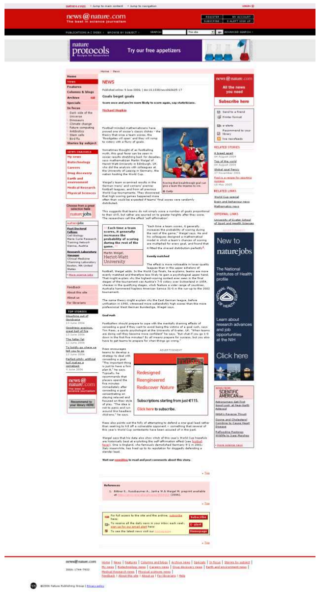
bioedonline.org



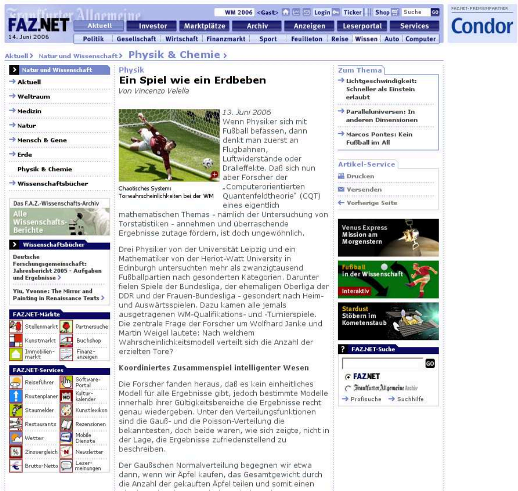
science.orf.at



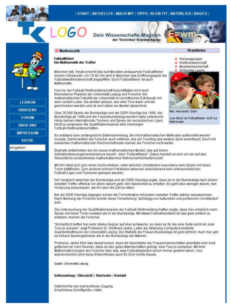
nature.com/news



faz.net



tk-logo.de



webindia123.com