"Traitor Tracing"

Boris Škorić

TU Eindhoven

Vakantiecursus *De Exacte Benadering*
24 Aug 2012

Overzicht

- Digitaal watermerken
- Coalitie-aanvallen
- Coalitie-resistente codes
- De Tardos code
  - eenvoudig te bewijzen eigenschappen
  - "exacte benaderingen"

Digitale watermerken

Doel

- Opsporen van "ongeauthoriseerde herdistributie" van audio/video
- Traceren van de bron = "traitor tracing"

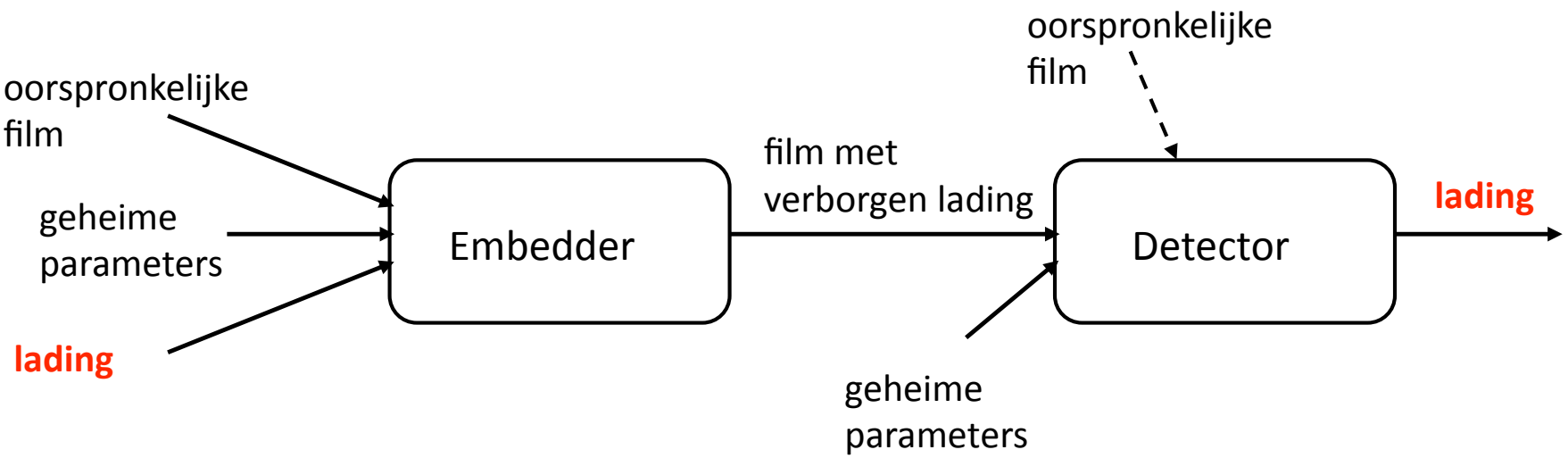
Methode

- Watermerk = geheime data verstopt in audio/video
  - mag luister & kijkplezier niet verstoren
  - moet eenvoudige manipulaties overleven (verandering van codec, helderheid etc)

Voorbeeld:

Watermerken worden gebruikt in Digitale Cinema

Forensisch watermerken



- Lading = unieke identificatie-code
- parameters verschillend voor elke film

Coalitie-aanvallen

"Coalitie van piraten"

- Aanvallers vergelijken wat ze ontvangen hebben
- De verschillen brengen watermerk aan het licht
- Op die plekken wordt watermerk gericht aangevallen

■ = "detecteerbare posities"

piraat #1	1	1	1	0	1	0	1	0	0	0	0	1
#2	1	0	1	0	1	0	1	0	1	0	1	1
#3	1	0	1	0	1	0	1	0	0	0	1	1
#4	1	1	1	0	0	0	1	1	0	0	0	1
resultaat na aanval	1	0/1	1	0	0/1	0	1	0/1	0/1	0	0/1	1

Coalitie-resistente watermerken

Verlanglijstje

- Bestand tegen $c \leq c_0$ aanvallers
- Zeer lage kans op valse beschuldigingen
- Lage kans om niemand te pakken
- ... en dat allemaal met weinig middelen!
  - 7bits/minuut video
  - niet te groot alfabet

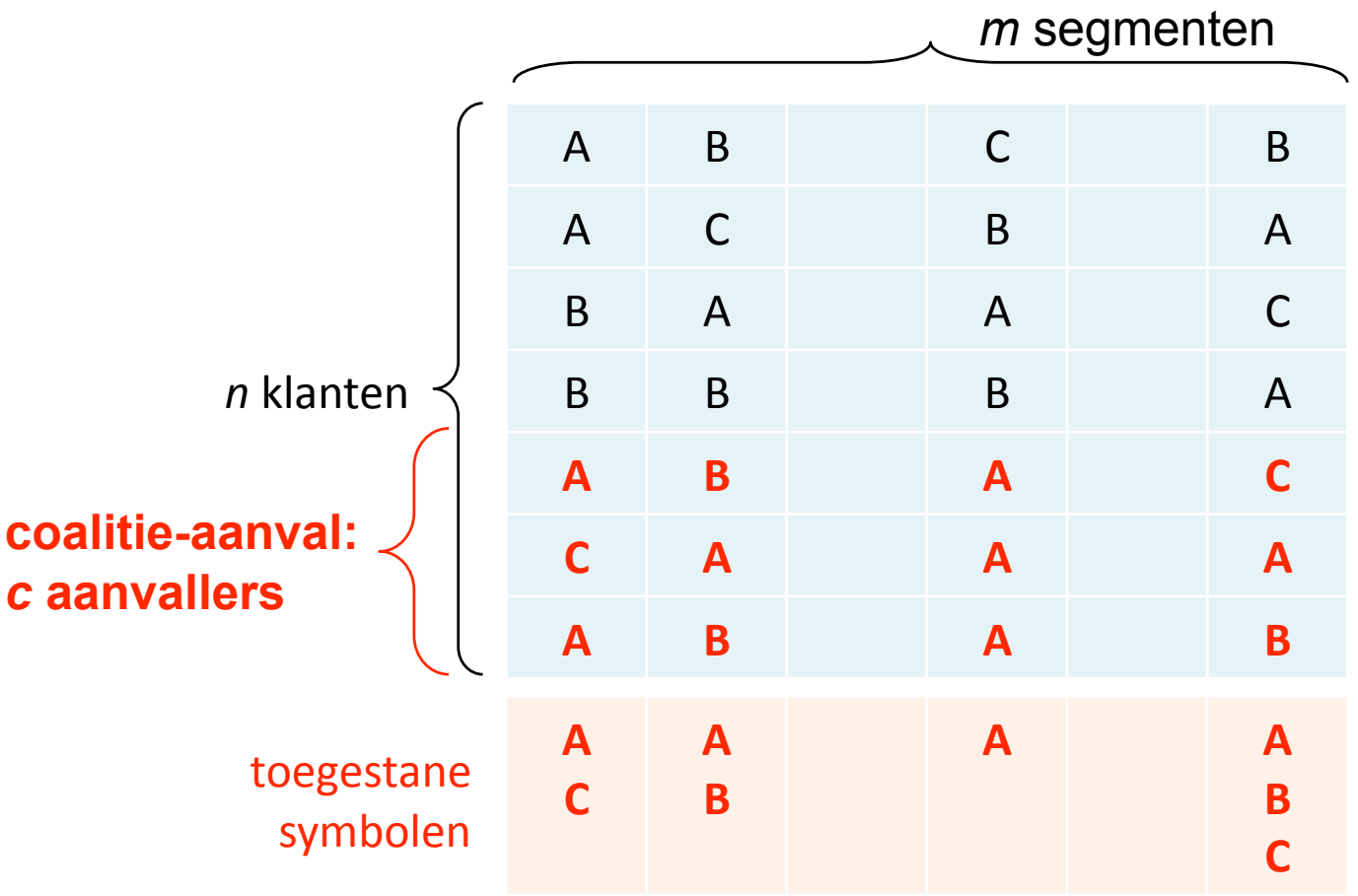
Wiskundig model

- Video opgedeeld in (abstracte) segmenten
  - uitgesmeerd in ruimte en tijd
- q-air alfabet
- Manipulatie alleen mogelijk in de detecteerbare posities
  - alleen keuze uit ontvangen symbolen
  - uitwissen kan niet
  - mengen kan niet

Aanvalsmodel

"restricted digit model"

Voorbeeld met q=3. Alfabet {A,B,C}



Geschiedenis: Lengte van de codes

Constructie

Boneh & Shaw 1998: $m = \mathcal{O}(c_0^4 \ln[n/\eta] \ln[1/\eta]), \quad q = 2$

Tardos 2003: $m = 100 c_0^2 \ln[n/\eta], \quad q = 2$

Chor et al 2000: $m = 4c_0^2 \log n, \quad q = 2c_0^2$

Staddon et al 2001: $m = c_0^2 \log_q(n), \quad q > m - 2$

Huang + Moulin; Amiri + Tardos 2009: $m = 2 c_0^2 \ln[n/\eta], \quad q = 2$

Boesten + Skoric 2011: $m = \frac{2}{q-1} c_0^2 \ln[n/\eta]$

Tardos 2003: $m = \Omega(c_0^2 \ln[n/\eta])$

Boneh and Shaw 1998: $m = \Omega(c_0 \ln[1/c_0\eta]), \quad q = 2$

Ondergrens

n = #klanten
m = codelengte = aantal segmenten
q = afmeting alfabet
η = Pr[er worden onschuldigen aangewezen]

De Tardos code

- Allereerste code met $m \propto c^2$ macht en klein alfabet
- In twee opzichten probabilistisch
  - staat kleine kans toe op valse beschuldiging en compleet missen van de aanvallers
  - constructie van de code is gerandomiseerd
- In 2003 verzonnen voor q=2, in 2007 uitgebreid naar algemene q

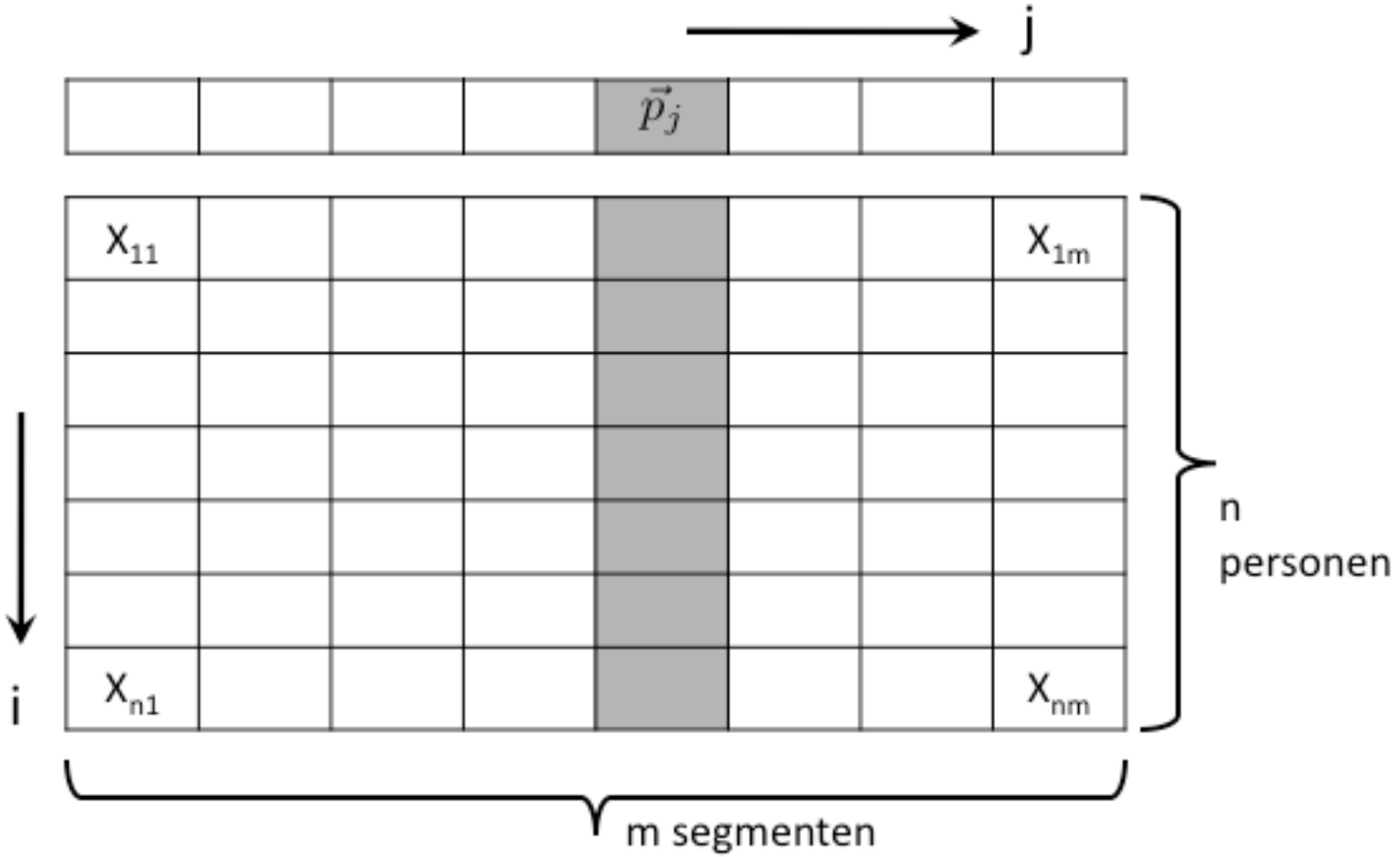
Tardos code: Het maken van de codewoorden

Onafhankelijk voor elk segment:

1. Kies "voorkeur"-vector ***p*** uit Dirichletverdeling
2. Genereer symbolen X aan de hand van ***p***

$$\boldsymbol{p}^{i} \sim F,\ F(\boldsymbol{p}) \propto \prod_{\alpha \in Q} p_{\alpha}^{-1+\kappa}$$

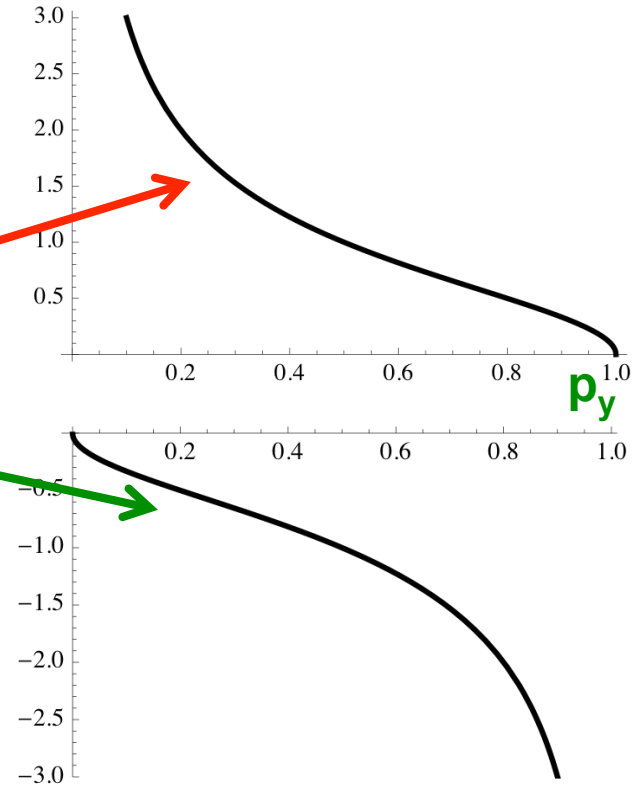
$$\Pr[X_{ij} = \alpha] = p_{\alpha}^{j}$$



Tardos code: Het traceren

- Er wordt een "ongeauthoriseerde" kopie gevonden
- Watermerk-detector ziet symbool y_j in segment j
- Reken voor elke klant *i* de "score" S_i uit
  - som van losse scores per segment: $S_i = \Sigma_j S_{ij}$

$$S_{ij} = \begin{cases} X_{ij} = y_j : & \sqrt{(1-p_y)/p_y} \\ X_{ij} \neq y_j : & -\sqrt{p_y/(1-p_y)} \end{cases}$$



- Klant *i* is verdacht als score S_i boven een bepaalde grens uitkomt

Tardos code: speciale eigenschappen

Scores van onschuldigen gedragen zich eenvoudig:

- Gemiddelde is nul in elk segment
- Variantie 1 in elk segment

$$\mathrm{E}[S_{ij}] = p_y \sqrt{\frac{1-p_y}{p_y}} - (1-p_y)\sqrt{\frac{p_y}{1-p_y}} = 0$$

$$\mathrm{E}[S_{ij}^2] = p_y \frac{1-p_y}{p_y} + (1-p_y)\frac{p_y}{1-p_y} = (1-p_y) + p_y = 1$$

- De Tardos score-functie is de enige met deze eigenschap
- Aanvallers hebben geen invloed op scores van onschuldigen

Speciale eigenschappen (II)

Wat kunnen we zeggen over de scores van de aanvallers?

- Definieer σ_α = #aanvallers dat symbool α ontvangt
- Totale score van de coalitie in een segment:

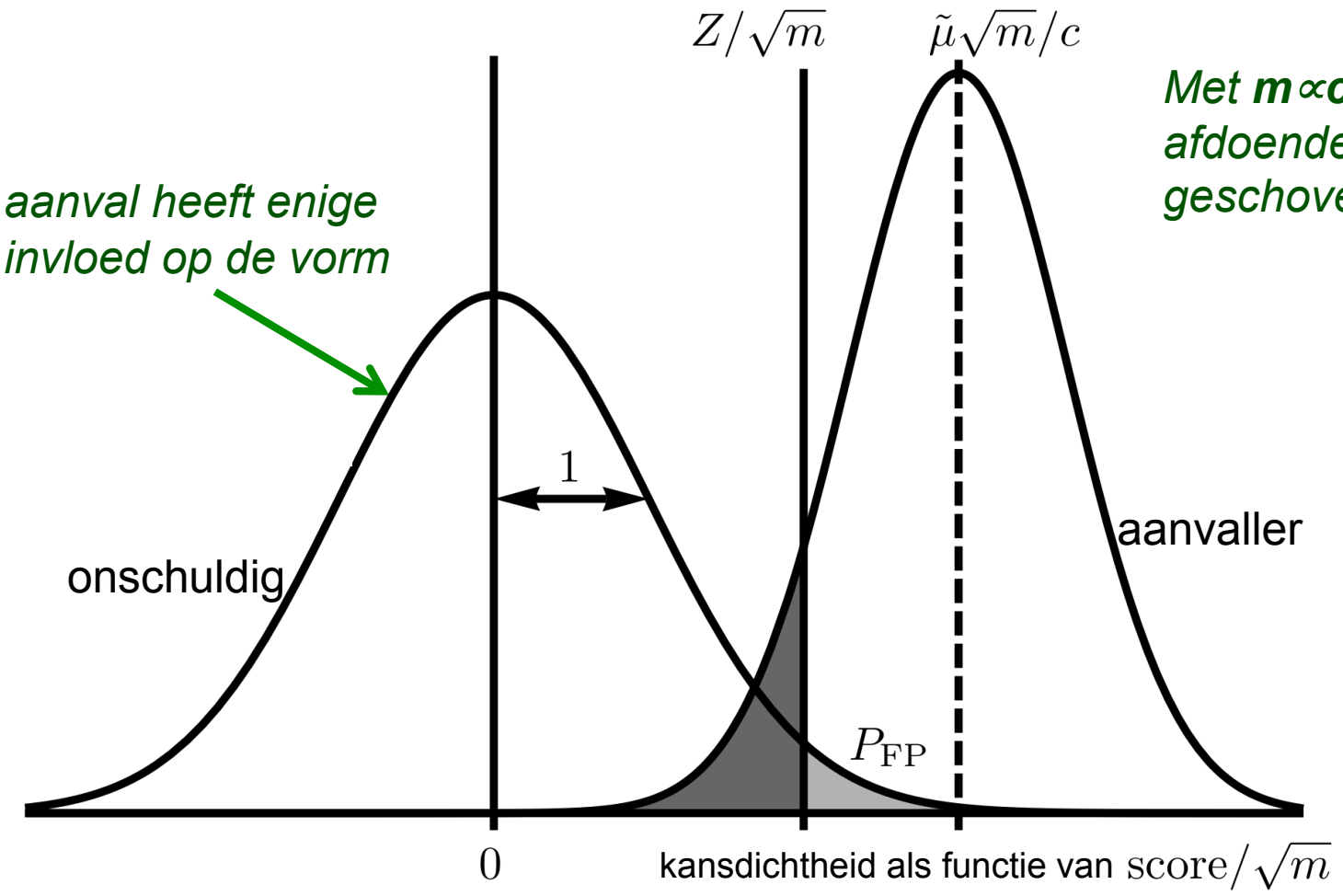
$$\sigma_y \sqrt{\frac{1-p_y}{p_y}} - (c-\sigma_y)\sqrt{\frac{p_y}{1-p_y}}$$

- Verwachtingswaarde over alle random beslissingen.
  (Kansverdeling van σ is multinomiaal.)

$$\mathrm{E}_{p\sigma\,y}[\text{tot.score}] = \mathrm{E}_p \sum_\sigma \binom{c}{\sigma} \mathrm{E}_y(\ldots) \frac{\mathrm{d}}{\mathrm{d}p_y}\left[p_y^{\sigma_y} (1-p_y)^{c-\sigma_y} \right]$$

- **Coalitie-score is geen stijgende functie van c.**
  Daalt zelfs een beetje.

Gevolgen van de speciale eigenschappen

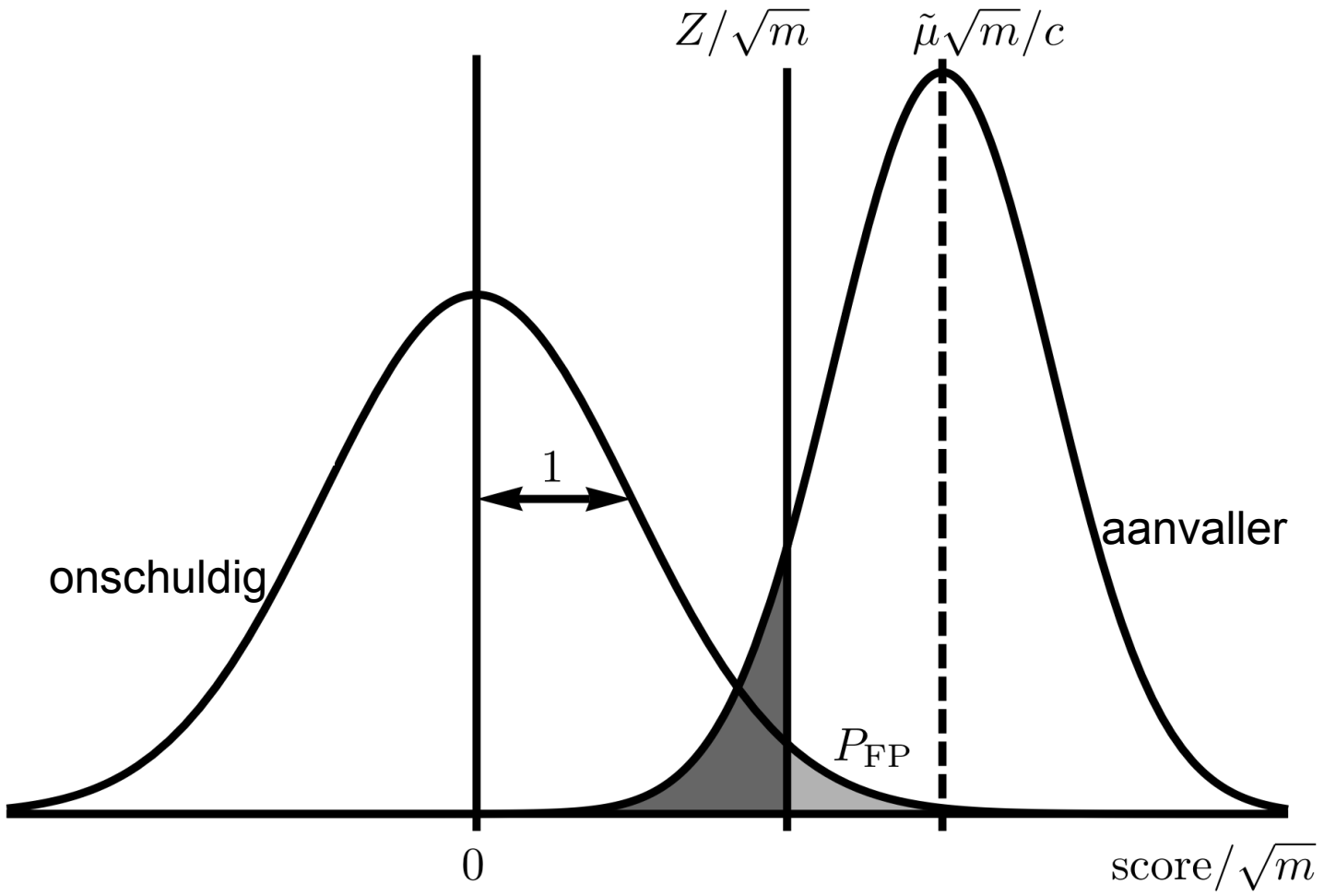


P_{FP} = kans dat (onschuldige) klant i onterecht wordt beschuldigd
m = #segmenten
c = #aanvallers
Z = grens

"No framing"

Wat als er meer aanvallers zijn dan geanticipeerd?

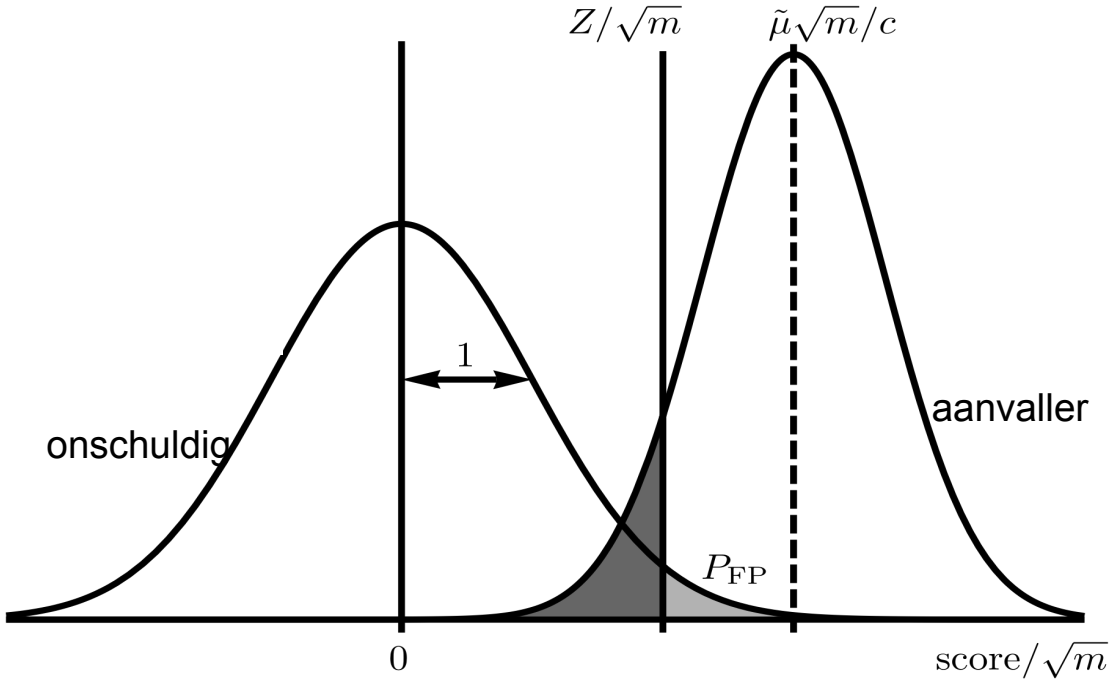
- Rechtercurve schuift links van de grens Z ⇒ **aanvallers niet gepakt**
- Linkercurve verandert nauwelijks ⇒ **geen onschuldigen gepakt**



Samenvattinkje tussendoor

- Constructie van de Tardos code en het score-systeem zijn heel eenvoudig
  - segmenten zijn onderling onafhankelijk
  - arbeid evenredig met *nm*
- Belangrijke eigenschappen zichtbaar met weinig analyse
  - de factor c^2 in $m = const.c^2 \ln(1/P_{FP})$
  - no framing
  - plaatsing grens $Z \propto \sqrt{m}$

Minder evidente eigenschappen

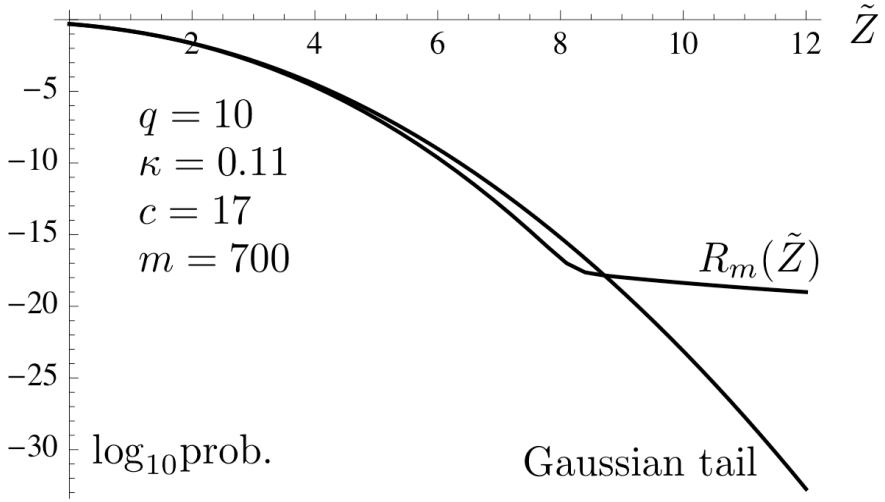


Centrale Limiet Stelling: kansverdelingen worden Gaussisch

- P_{FP} wordt Complementaire Errorfunctie Erfc
- gebruik $[\mathrm{Erfc}^{\mathrm{inv}}(x/\sqrt{2})]^2 < \ln(\sqrt{\pi}/x)$
- na enig gepruts volgt de factor $\ln(1/P_{FP})$ in $m = const.c^2 \ln(1/P_{FP})$

Nog veel minder evidente zaken

- De precieze waarde van de constante in m = *const*.c^2 ln($1/P_{FP}$)
  - In praktische situaties *const*<9
  - Asymptotisch voor grote c: *const* → 2/(q-1).
    Vereist ingewikkeld score-systeem.
- De exacte vorm van de kansverdelingen
  - convolutie-stelling voor "optellen" van kansverdelingen in het Fourier-domein: als **Z=X+Y**, X~f, Y~g, en Z~h , dan *F*[h] = *F*[f] *F*[g]
  - "exacte benadering" door Taylor expansie in $m^{-1/2}$



Samengevat

- Kat-en-muis-spel van aanval en verdediging bij watermerken leidt tot flink wat wiskunde
  - Coderingstheorie
  - Informatietheorie
  - Statistiek
  - Functionaal-analyse
  - Reeksen
- We hebben de Tardos code in detail bekeken
  - optimaal tegen grote coalities
  - no framing
  - simpel in gebruik
  - makkelijk te begrijpen