

Information- and Coding theory in biometrics

May 24

Kruger Park, SA

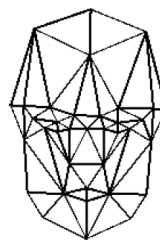
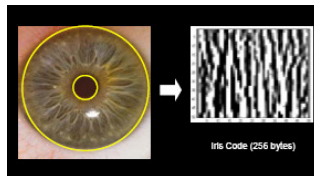
A.J. Han Vinck



in cooperation with V. Balakirsky

University Duisburg-Essen
Germany
May 2010

Goal: use biometrical features as passwords



biometrics do change

Example 1



Example 2

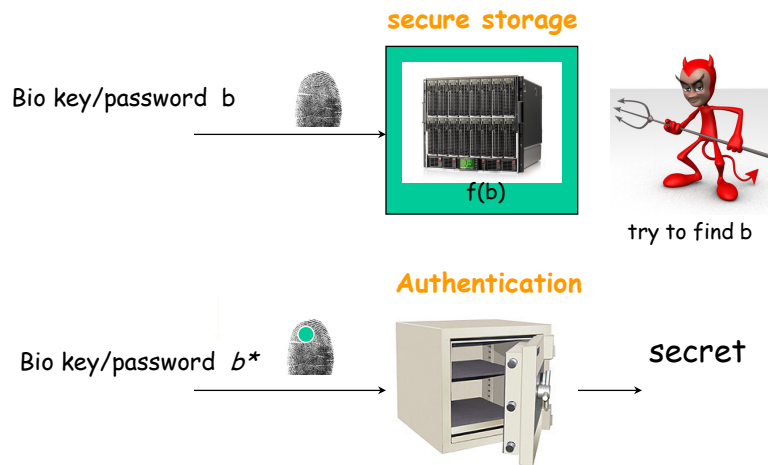


6/7/2010

A.J. Han Vinck

3

Problem: secure storage and biometric authentication

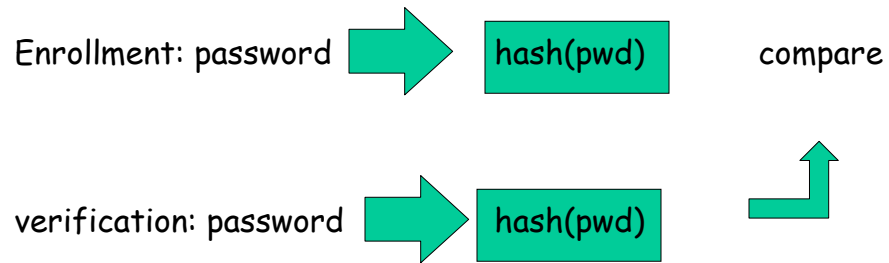


6/7/2010

A.J. Han Vinck

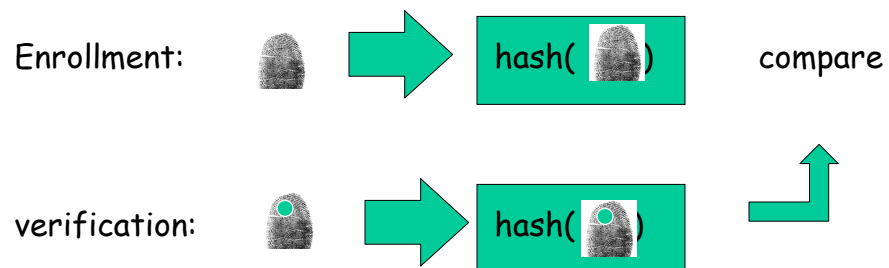
4

Illustration of the problem



5

Illustration of the problem



6

hash functions of biometrics can not be used as passwords

for a vector c and a noisy version $c' = c \oplus \text{noise}$

hash property: $\text{hash}(c' \approx c) \neq \text{hash}(c)$
single error $\Rightarrow n/2$ differences

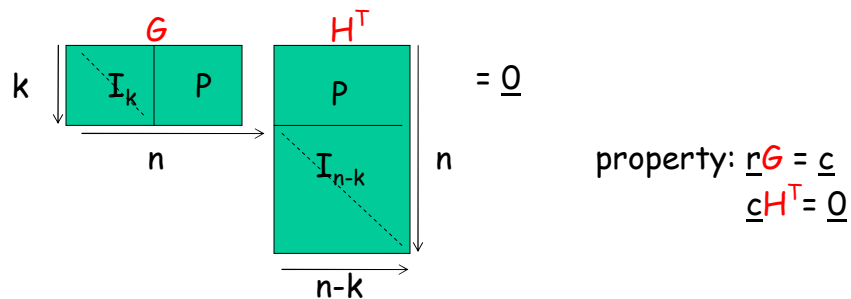
may be we can use Error-correction:

$\text{dec}(c' \approx c) = \text{dec}(c)$
equality for $2t < d_{\min}$

7

idea: Use redundancy to correct errors in the Bio

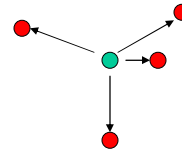
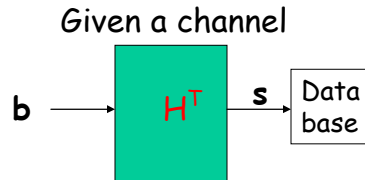
Properties of a linear code: length n , k information digits
odd minimum distance d_{\min}



Property: let $e_1 H^T = s_1$ and $e_2 H^T = s_2$; $e_1 \neq e_2$

then $s_1 \neq s_2$ for $|e_1|$ and $|e_2| < d_{\min} / 2$ because...

Maximum A posteriori Probability (MAP) receiver (minimum error probability)



Attacker of DB: for every s , guess a particular b_i

- the best guess is the b_i for which $P(b_i \text{ stored as } s | s)$ is maximum

$$P(\text{correct} | s) = \max_b P(b | s)$$

$$\bar{P}(\text{correct}) = \sum_s P(s) \max_b P(b | s) = \sum_s \max_b P(s | b) P(b) \quad \text{Bayes rule}$$

What about the security of stored information

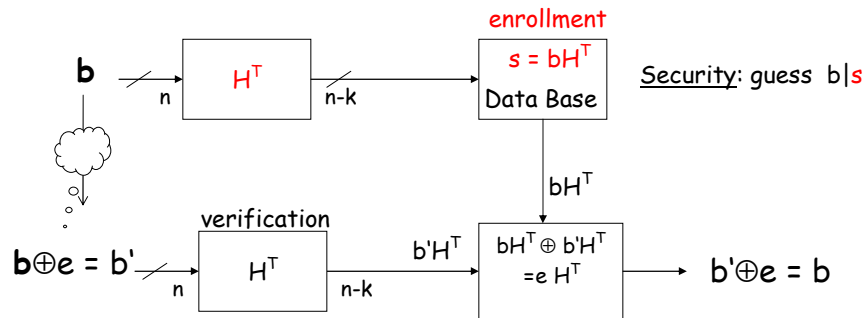
$s = bH^T$ THUS: there are 2^k „candidate“ fingerprints

- A. given $s = bH^T$, the entropy $H(b|s) \leq k$
- B. $H(s) + H(b|s) = H(b) + H(s|b) = H(b)$ since b determines s
 $H(b|s) = H(b) - H(s) \geq H(b) - (n-k)$ ENTROPY LOSS

(# possible candidates reduced by factor of $2^{(n-k)}$)

- For k small: $H(b|s) \Rightarrow 0$ bad security
- For k large: $H(b|s) \Rightarrow H(b)$ good security

construct b from a noisy version b' and syndrome s



Conclusion:

- For k small: good reconstruction, bad security
- For k large: bad reconstruction, good security

Example: BCH codes (bits) test for a valid syndrome

For binary BCH codes: $n = 256$, $k = 224$ bits, $d_{\min} = 7$

- False Rejection Rate = $P(\#errors \geq 4) \approx (100p)^4$;
too many differences
- False Acceptance Rate $< 2^{-8}$
random vector insided decoding region
- Security: 2^{-224}

for t-error correction: test whether we have a valid syndrome

BCH: redundancy $n-k \sim t \log_2 n$

$$\text{FRR} \Rightarrow \binom{n}{t+1} p^{t+1} \approx (np)^{t+1}; \quad \text{FAR} \Rightarrow \binom{n}{t} / 2^{t \log_2 n} \approx 1; \quad \text{SEC} \Rightarrow 2^{-k}$$

P(too many errors)

P(inside decoding region)

guess

BCH: $n^* - k^* = n - k \sim t \log_2 n$



$$\text{FRR} \Rightarrow \binom{n^*}{t+1} p^{t+1} \approx (n^* p)^{t+1}; \quad \text{FAR} \Rightarrow \left(\frac{n^*}{n}\right)^t < 1; \quad \text{SEC} \Rightarrow 2^{-k^*}$$

decrease

decrease

increase

6/7/2010

A.J. Han Vinck

option: reduce decoding region

BCH: redundancy $n-k \sim t \log_2 n$

$$\text{FRR} \Rightarrow \binom{n}{t+1} p^{t+1} \approx (np)^{t+1}; \quad \text{FAR} \Rightarrow \binom{n}{t} / 2^{t \log_2 n} \approx n^{t-t}; \quad \text{SEC} \Rightarrow 2^{-k}$$

P(too many errors)

P(inside decoding region)

guess

decode less than possible



$$\text{FRR} \Rightarrow \binom{n}{t^*+1} p^{t^*+1} \approx (np)^{t^*+1}; \quad \text{FAR} \Rightarrow \binom{n}{t^*} / 2^{t^* \log_2 n} \approx n^{t^*-t^*}; \quad \text{SEC} \Rightarrow 2^{-k}$$

increase

decrease

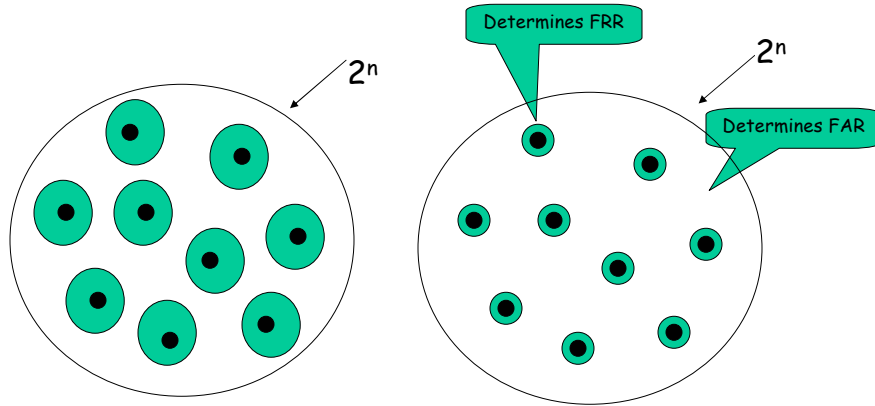
same

6/7/2010

A.J. Han Vinck

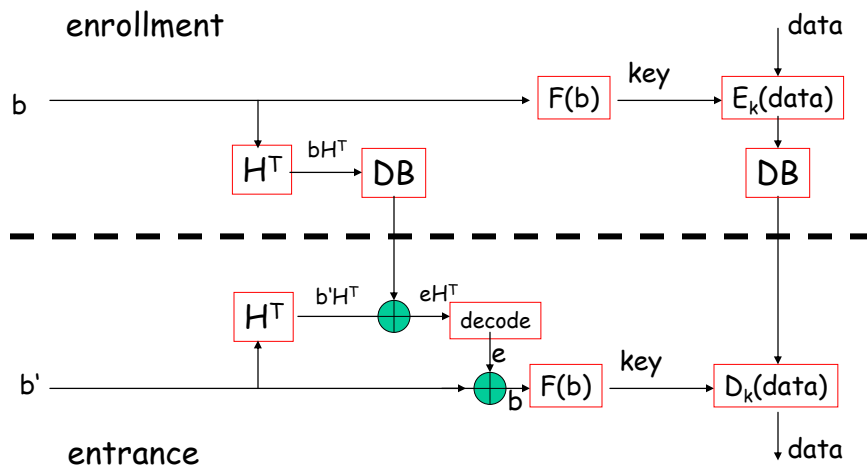
14

As a picture

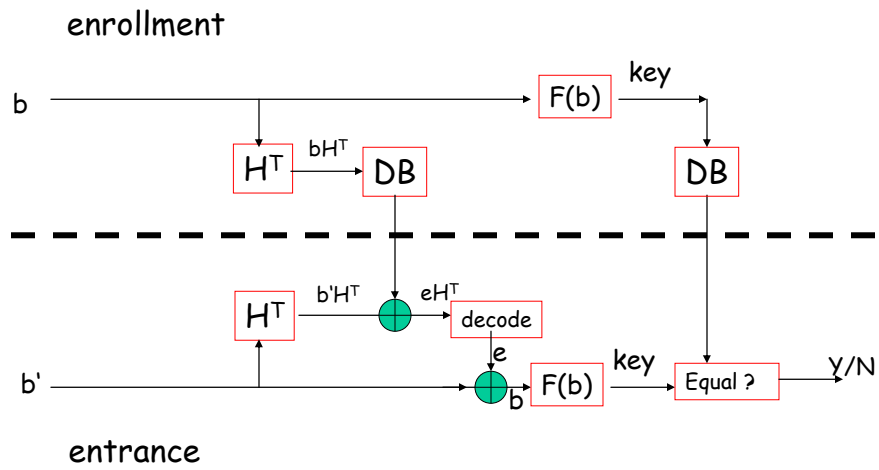


Number of codewords and length stays the same

It is time for an application



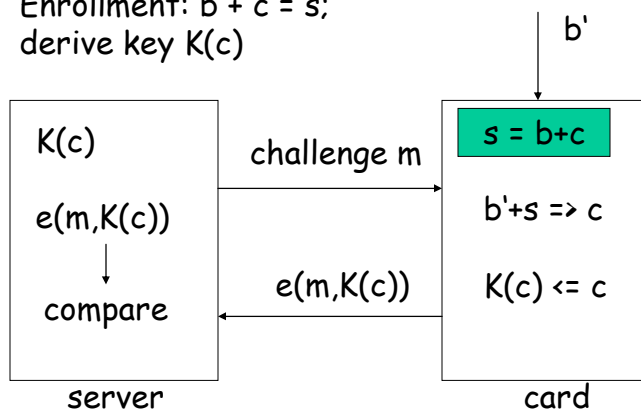
Another application



17

Challenge response

Enrollment: $b + c = s$;
derive key $K(c)$



18

Fuzzy vault

lock

Key = b

secret

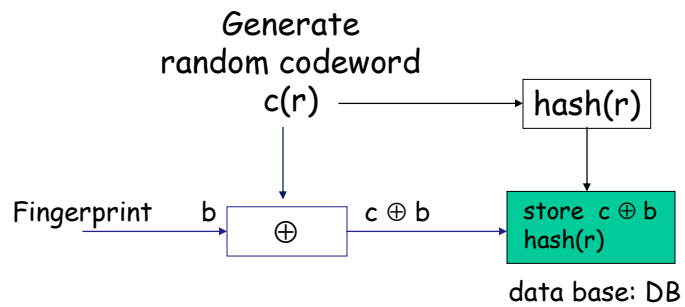
unlock

Key = b'

secret

19

Another scheme: Enrollment

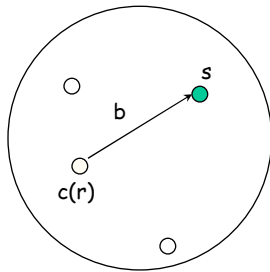


Condition: given $c \oplus b$ and $\text{hash}(r)$
it is hard to estimate b or $c(r)$

Idea: Juels-Wattenberg

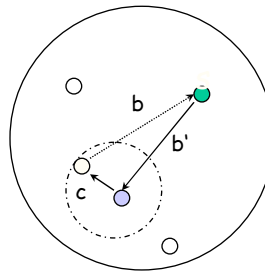


Enrollment: $b = \text{fingerprint}$



- 2^k Codewords c
- choose random r
- store $s : s = c \oplus b$

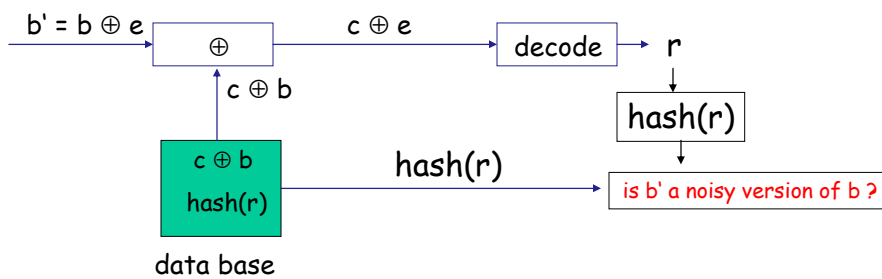
Secure sketch: input b'



- decode c from $s \oplus b'$
- calculate $s \oplus c = b$

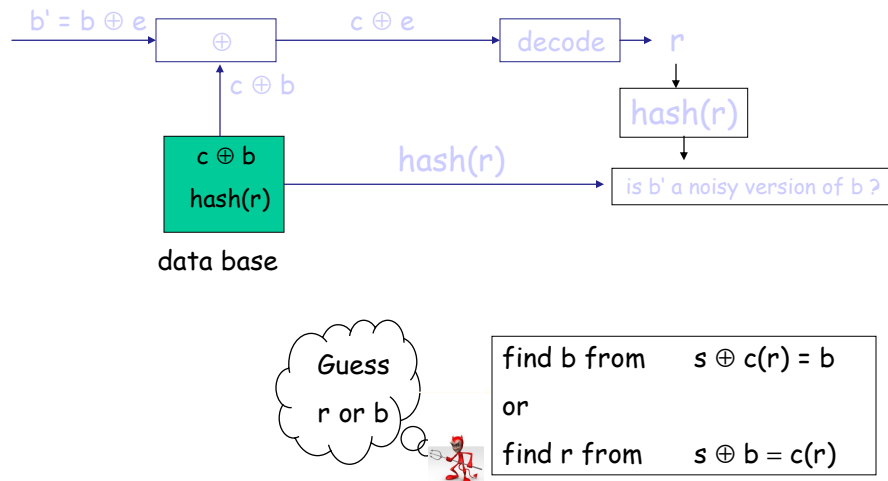
21

authentication



FRR: valid b' rejected; FAR: invalid b' accepted;

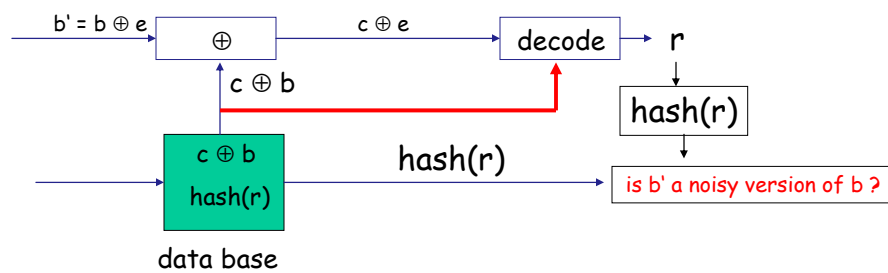
attacker



Han Vinck

23

Improved legal detector

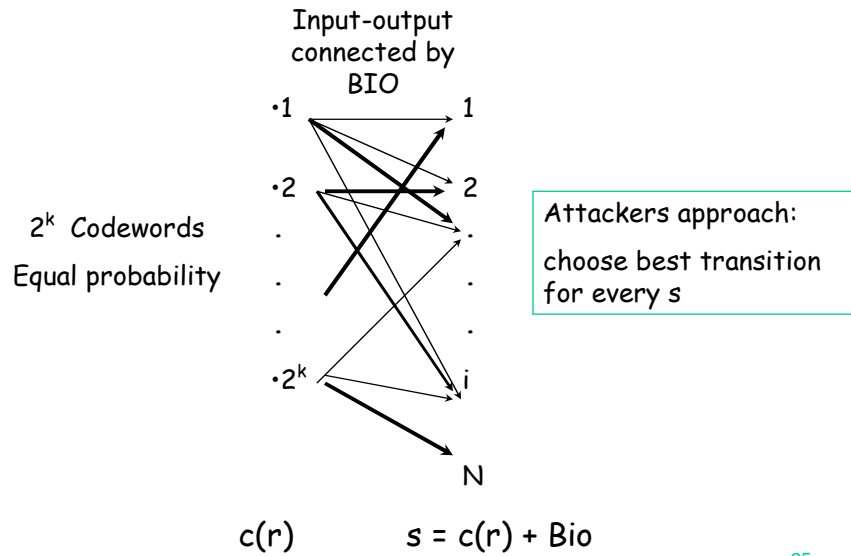


FRR: valid b' rejected; FAR: invalid b' accepted;

Han Vinck

24

Authentication as a channel



25

Attacker: minimum error probability estimator

- attacker information: $s = c(r) + b$

$$\begin{aligned}
 P_{\text{attack}}(\text{correct}) &= \sum_s \max_c P(s|c)P(c) \\
 &= \sum_s \frac{1}{2^k} \max_c P_{\text{bio}}(c+b|c) \\
 &\leq 2^{n-k} \max P_{\text{bio}}(b) \quad \text{we loose a factor of } 2^{n-k}
 \end{aligned}$$

Alternative:

Guess r directly. Probability of success $\geq 2^{-k}$

Guess b directly. Probability of success $\max P_{\text{bio}}(b) \geq 2^{-H(b)}$

remember the entropy loss of (n-k) bits?

26

Performance: minimum error probability estimator

legal receiver: $b' + c + b = c + e$

$$P_{\text{legal}}(\text{correct}) = \sum_{s=c+e} \max_c P(c+e|c)P(c) = \sum_{s=c+e} \frac{1}{2^k} \max_c P(c+e|c)$$

P depends on the error vector and the code design

27

Improved receiver uses : $s = c + e$ and $z = c + b$

$$P_{\text{improved}}(\text{correct}) = \sum_{s,z} \max_c \frac{1}{M} \{P(c+e|c)P_{\text{bio}}(c+b|c)\}$$
$$\approx 2^{nh(p)} 2^{nH(b)} \max_c \{P(c+e|c)P_{\text{bio}}(c+b|c)\}$$

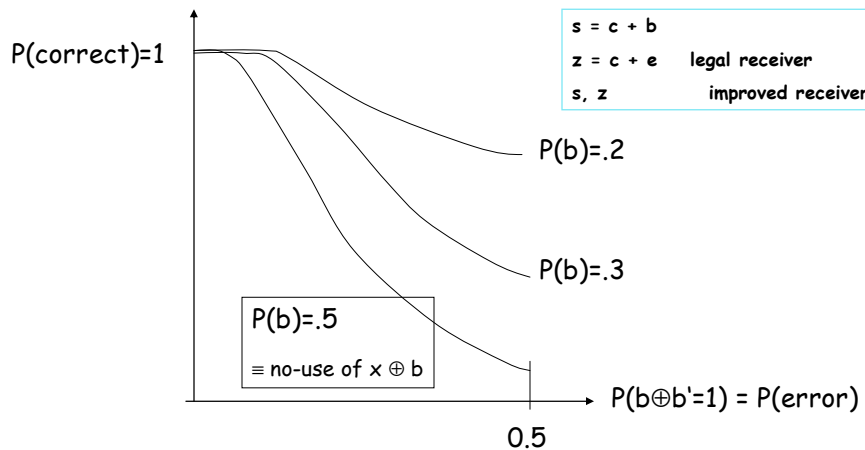
Example: $P(e=1) = p$; $P(b=1) = q$

then: $\max P(\underline{e})P(\underline{b}) = \max \{p/(1-p)\}^i * \{q/(1-q)\}^j$

For $q = \frac{1}{2}$ we have no gain!

28

Example: BCH (15,5)



29

Using the Fano inequality

- $P(\text{incorrect guess}) := P_{\text{inc}} ; M = 2^k$.

Fano inequality: $H(c,s) = H(s) + H(c|s) = H(c) + H(s = b+c|c)$

- $H(c|s) = H(c) + H(b) - H(s) \leq h(P_{\text{inc}}) + P_{\text{inc}} \log_2 (M-1) \leq 1 + k P_{\text{inc}}$

- From this: $P(\text{correct}) = 1 - P_{\text{inc}} \leq (n + 1 - H(b)) / k$
 $\leq (n + 1 + \log_2 \max P_{\text{bio}}) / k$

Conclusion:

low probability of successful attack for a small $\max P_{\text{bio}}$ and large k .

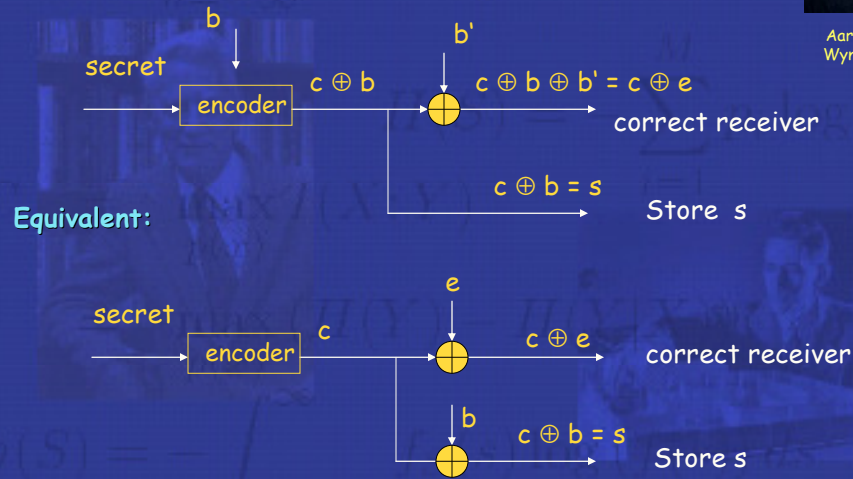
30



equivalent noisy wiretap channel model



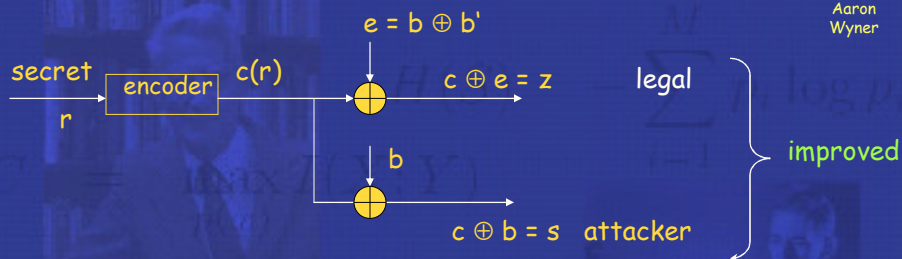
Aaron Wyner



Equivalent: noisy wiretap channel model



Aaron Wyner



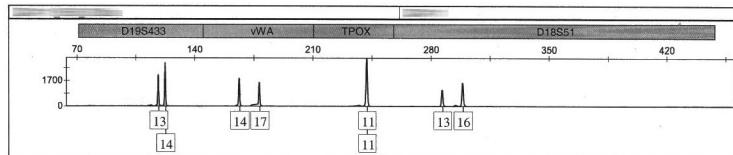
secrets transmitted: 2^{nC_s} where C_s is the secrecy capacity

secrecy capacity for the wiretap channel: $= H(b) - H(e)$

secrecy capacity improved: $= H(b) - H(e) + H(s|z) - H(s|x)$

Application for DNA

- set of alleles: D8S1179, VWA, FGA, TH01, ...



- every allele is characterized by 1 or 2 numbers from a set
 - i.e.; TH01 = { 6,7,8,9 }
- probability distribution of the numbers
 - $P(6,7,8,9) = (0.23, 0.19, 0.09, 0.49)$
 - i.e. $P(6) = P(6,6) = 0.0529$; $P(6,9) = 0.2254$; ...
- We can calculate: entropy, FAR, FRR, EER, etc.

(some) errors in the allele information

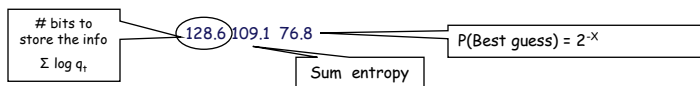
• ERROR TYPES

- Drop in (10) := (10,12)
- Drop out (10,12) := (10)
- Allelic shift (10,12) := (10,13)

- Probability of error: ~ 0.05 %

Table of allele information

t	Name	log q _t	H _t	G _t	t	Name	log q _t	H _t	G _t
15	D5S818	3.91	3.11	1.81	1	D8S1179	4.39	4.08	3.01
16	TPOX	3.91	2.91	1.79	2	D3S1358	3.91	3.71	2.87
17	CFIPO	3.91	3.16	2.16	3	VWA	4.39	4.13	3.12
18	D8S1179	5.49	4.49	3.15	4	D7S820	4.39	4.07	3.25
19	VWA-1	4.39	4.13	3.12	5	ACTBP2	7.71	7.43	6.13
20	PentaD	5.17	4.32	3.13	6	D7S820	4.81	4.24	3.31
21	PentaE	6.91	5.87	4.02	7	FGA	5.49	4.92	3.54
22	DYS390	4.39	3.24	2.06	8	D21S11	4.81	4.13	3.01
23	DYS429	3.91	2.97	1.78	9	D18S51	5.78	5.28	4.43
24	DYS437	2.58	2.26	1.58	10	D19S433	4.39	3.59	2.33
25	DYS391	3.32	1.90	1.11	11	D13S317	4.81	4.15	2.56
26	DYS385	5.17	3.61	1.72	12	TH01	3.32	2.85	2.07 = $-\log_2 P(\max)$
27	DYS389I	2.58	2.01	1.18	13	D2S138	6.04	5.60	4.23
28	DYS389II	3.91	3.14	2.04	14	D16S539	4.81	3.78	2.25



35




University Duisburg-Essen

digital communications group



Running content

- Juels-Wattenberg scheme
- Juels-Sudan (Reed Solomon-based) 
- Dodis Equal-weight codes and permutations

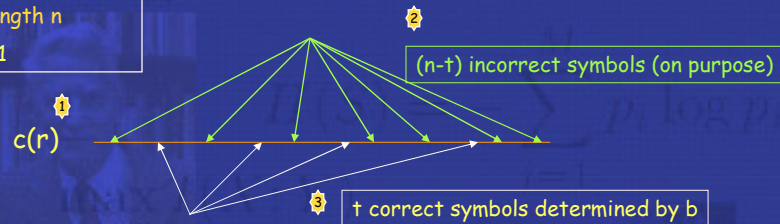
$$h(S) = \sum_{s \in S} p(s) \log_2 \frac{1}{p(s)}$$



Reed Solomon encoding Juels-Sudan (idea)



Secret: k symbols
 Codeword: length n
 $d_{min} = n - k + 1$



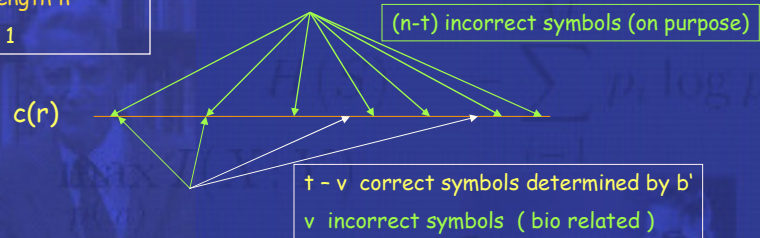
security: $c(r)$ contains $(n-t)$ errors:
 if $2(n-t) + 1 > d_{min} = n - k + 1$
 then r „not decodable“ from $c(r)$



Reed Solomon decoding (idea)



Secret: k q -ary symbols
 Codeword: length n
 $d_{min} = n - k + 1$



authentication: use t positions determined by b'
 if $2v + 1 \leq d'_{min} = t - k + 1$ (d_{min} reduced by known error positions)
 then r „decodable“ from $c(r)$



Performance: minimum error probability estimator

security:

for bio with t q -ary symbols $2(n-t)+1 > n-k+1 \Rightarrow 2t < n+k$

for v error correction: $2v \leq t-k \Rightarrow 2v < n-t$

REMARK: for $t \leq n/2$ we point at incorrect positions!



Security: guess codeword

Simple attack: guess k correct positions in c

$$\text{prob(correct)} = \frac{\text{\# of correct guesses}}{\text{total \# of guesses}} = \frac{\binom{t}{k}}{\binom{n}{k}} \rightarrow \left(\frac{t}{n}\right)^k$$

For $n = q$, guessing the correct r

without c

$$P_{\text{correct}} = q^{-k}$$

given c

$$P_{\text{correct}} \approx t^k q^{-k}$$

Note: for $t = k$, no loss



Security: guess bio

MAP decoder: \hat{y} is the result that follows from codeword c and Bio b

$$\begin{aligned}
 P_{\text{correct}} &= \sum_y \max_c P(y | c) P(c) \\
 &\leq q^n q^{-k} \max_c P(y | c) \\
 &\leq q^{n-k} \max_{\text{bio}} P(b) q^{-(n-t)} \\
 &= q^{t-k} \max_{\text{bio}} P(b)
 \end{aligned}$$

guessing the correct b given c is worse with factor $q^{t-k} = q^{2r}$



Comparison (with errors)



Assume: # errors = ϵ ; $n-k = 2\epsilon$ $|b| = q^t$;

• Juels-Wattenberg (q -ary):



* probability (Correct guess) = $q^{n-k} / |b| = q^{2\epsilon-t}$
 $n - k = 2\epsilon$; entropy loss 2ϵ

Juels-Sudan:



* probability (Correct guess) = $t^k q^{2\epsilon-t}$
 $t - k = 2\epsilon$; entropy loss $2\epsilon + k \log_q t$

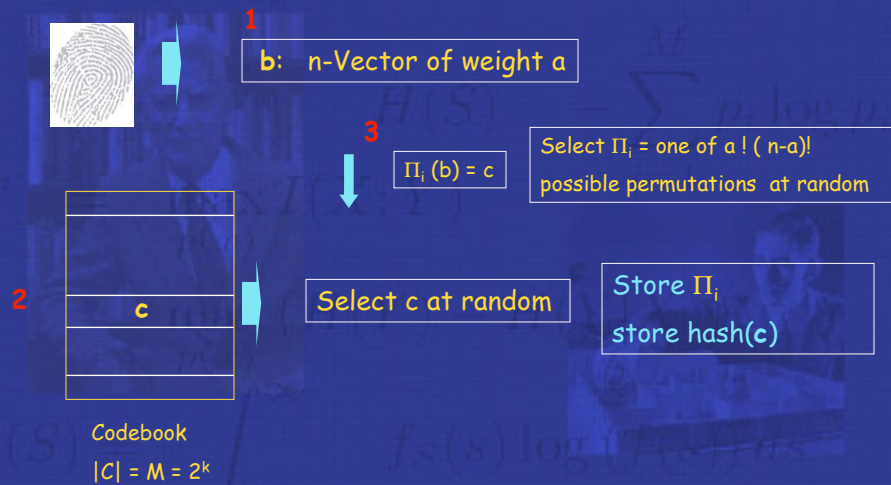


Running content

- Juels-Wattenberg scheme
- Juels-Sudan (Reed Solomon-based)
- Dodis Equal-weight codes and permutations



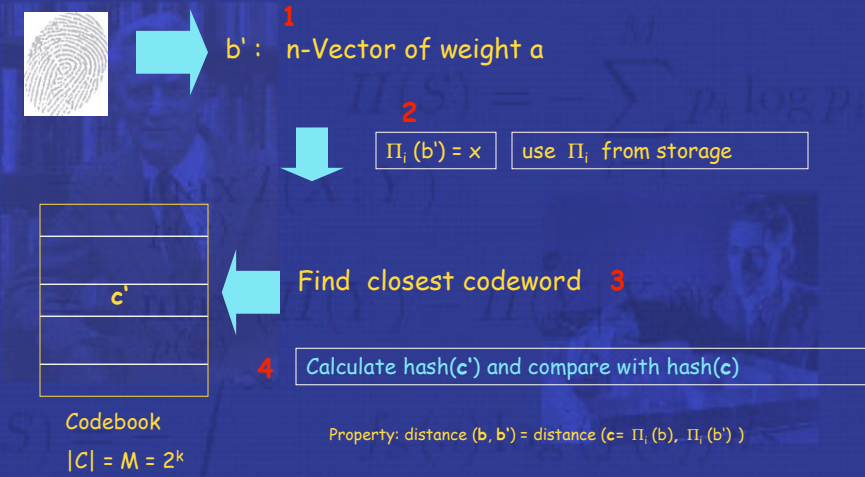
Dodis, Reyzin and Smith enrollment



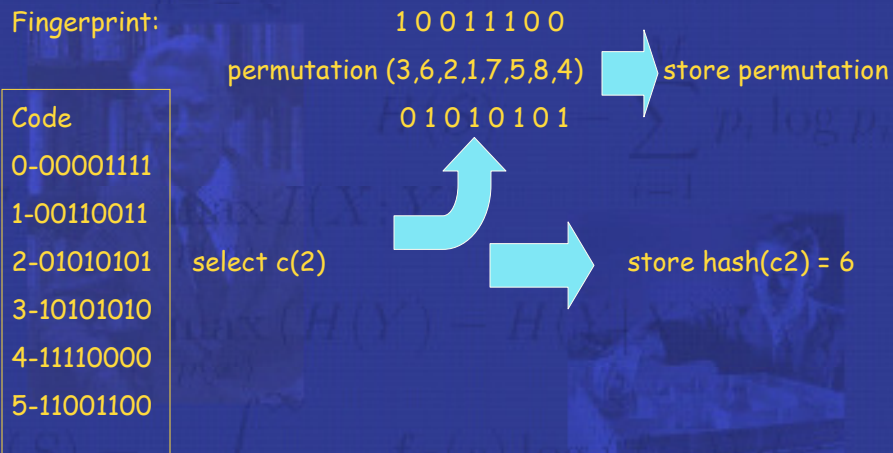


Dodis, Reyzin and Smith authentication

input: hash(c), permutation



example





example

False Fingerprint:

permutation (3,6,2,1,7,5,8,4) →

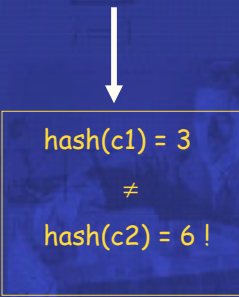
10110001

10010011

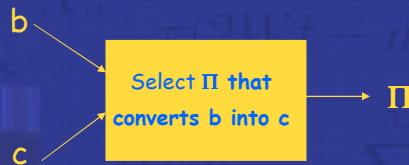
decode:

00110011 c(1)

Code
0-00001111
1-00110011
2-01010101
3-10101010
4-11110000
5-11001100



attacker (MAP): search for valid b



Attacker:

go through all possible c and output b that maximizes $\Pr(b = \Pi^{-1}(c))$

$$\begin{aligned} \bar{P}_{\text{correct}} &= \sum_{\pi} \max_{c,b} \Pr(\pi | c, b = \pi(c)) \Pr(c, b = \pi(c)) \\ &= \sum_{\pi} \frac{1}{a!(n-a)!} \max_{c,b} \Pr(c, b = \pi(c)) = \sum_{\pi} \frac{1}{a!(n-a)!} \max_{c,b} \Pr(c) \Pr(b = \pi(c) | c) \\ &= \sum_{\pi} \frac{1}{a!(n-a)!} \frac{1}{2^k} \max_{c,b} \Pr(b = \pi(c) | c) \leq \binom{n}{a} 2^{-k} \max_{b \in \text{bio}} \Pr(b) \\ &\approx 2^{nh(a/n) - k} \max_{b \in \text{bio}} \Pr(b) \end{aligned}$$

Redundancy = $nh(a/n) - k$



Actual research projects

- Translation from bio to binary needed
 - Problem of „distance“
- estimating statistical properties of b
- security for non-uniform b
- errors are „a-typical“ (shifts, drop-out, etc)



Conclusion and problems

Error correcting codes make Biometric Authentication possible

Price for recovery:

$$P_{\text{attack}}(\text{success}) \leq q^{\text{redundancy}} \max_{b \in \text{bio}} \Pr(b)$$

direct guess

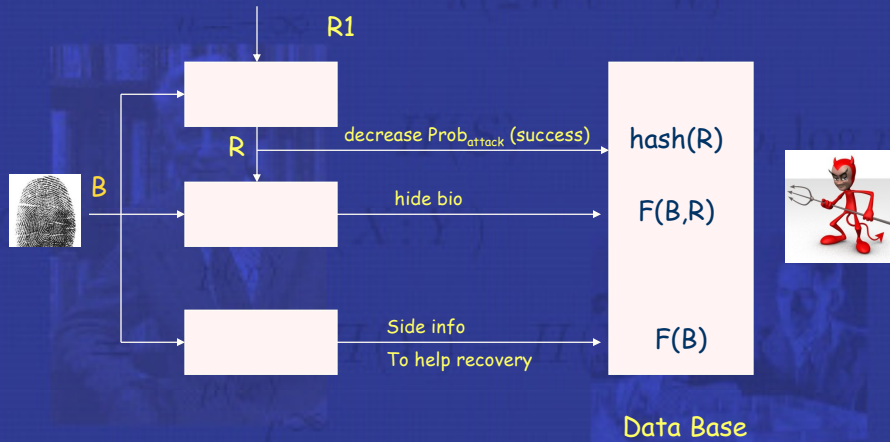
$$* \text{ guess } k \text{ symbols of } r \rightarrow P_{\text{attack}}(\text{success}) \geq q^{-k}$$

$$* \text{ guess bio} \rightarrow P_{\text{attack}}(\text{success}) \geq \max_{b \in \text{bio}} \Pr(b)$$

- Average probability of success for an attacker increases



Improvement at enrollment



"An Achievable Region for the Gaussian Wiretap Channel with Side Information,"
IEEE Transactions on Information Theory, May 2006,
C. Mitrpant, A.J. Han Vinck and Yuan Luo

