Οικονομικό Πανεπιστήμιο Αθηνών
Τμήμα Πληροφορικής
ΠΜΣ

Κρυπτογραφία και Εφαρμογές

Μαριάς Ιωάννης
marias@aueb.gr

Μαρκάκης Ευάγγελος
markakis@gmail.com

Περίληψη

- Shannon theory
  - ✓ Εντροπία
  - ✓ Μελέτη κρυπτοσυστημάτων με θεωρία πληροφορίας
  - ✓ Τέλεια μυστικότητα
  - ✓ Πλεονασμός – Περίσσεια γλώσσας
- Ανίχνευση γλώσσας
  - ✓ Έλεγχος Κάπα
  - ✓ Έλεγχος Χι

Θεωρία Πληροφορίας

C. E. Shannon, Communication theory of secrecy systems, *Bell Systems Technical Journal*, 656-715, 1949

- ✓ Cryptography via information theory

- ✓ Έστω X τυχαία μεταβλητή που παίρνει τιμές από ένα πεπερασμένο σύνολο $\{x_1, x_2, \ldots x_n\}$ με πιθανότητες $P(X = x_i) = p(x_i)$ όπου $0 \leq p(x_i) \leq 1$ για $i = 1, \ldots, n$, και $\Sigma_i\, p(x_i) = 1$
- ✓ Αν Y επίσης τυχαία μεταβλητή, η απο κοινού κατανομή θα συμβολίζεται με $p(x_i, y_j) = P(X=x_i, Y=y_j)$
- ✓ Αν X, Y είναι ανεξάρτητες τυχαίες μεταβλητές τότε

$$p(x_i, y_j) = p(x_i)\, p(y_j), \text{ για κάθε } x_i, y_j$$

Δεσμευμένες πιθανότητες

- ✓ $p(x_i|y_j)$ = πιθ/τα η X να πάρει την τιμή x_i δεδομένου ότι η Y έχει την τιμή y_j
- ✓ $p(x_i|y_j) = p(x_i, y_j) / p(y_j)$, όταν $p(y_j) > 0$
- ✓ Αν X, Y είναι ανεξάρτητες τυχαίες μεταβλητές τότε

  $p(x_i|y_j) = p(x_i)$

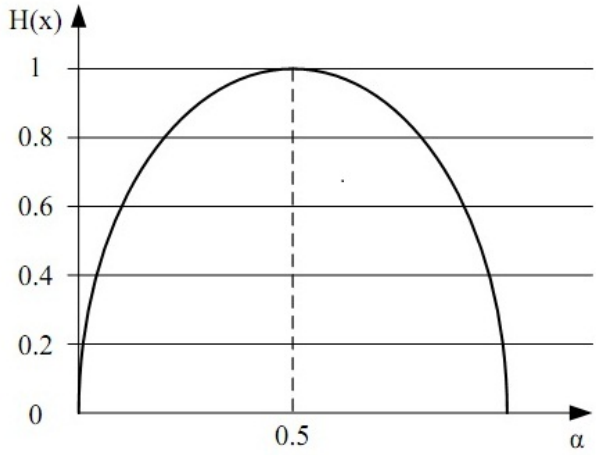
Θεωρία Πληροφορίας

- ✓ Εντροπία: μαθηματική απεικόνιση για το μέγεθος της πληροφορίας που παρέχεται από παρατήρηση της X
  - Ή η αβεβαιότητα για την έκβαση πριν από την παρατήρηση της X
- ✓ Ορισμός εντροπίας (Πληροφορία κατά Shannon): Η ποσότητα της αβεβαιότητας μίας μεταβλητής X είναι:

$$\mathbf{H(X) = -\Sigma_i\, p(x_i) \log p(x_i) = \Sigma_i\, p(x_i) \log(1/p(x_i))}$$

- ✓ Ιδιότητες:
  - $H(X) = 0$ iff $p(x_i) = 1$ για κάποιο i, και $p(x_j) = 0$ για κάθε $j \neq i$
    - Δεν υπάρχει αβεβαιότητα για το αποτέλεσμα
  - $H(X) = \log n$ iff $p(x_i) = 1/n$ για κάθε i, $1 \leq i \leq n$
    - Και οι n εκβάσεις, $x_1,\ldots,x_n$, ειναι ισοπίθανες

- ✓ Παράδειγμα: έστω τυχαία μεταβλητή X με σύνολο τιμών {0,1} και p(0) = a, p(1) = 1-a.
- ✓ H(X) = -aloga -(1-a)log(1-a)
- ✓ Αν a=0 ή a=1, τότε H(X)=0
  - Δεν υπάρχει αβεβαιότητα για το αποτέλεσμα
- ✓ Για κάθε άλλη τιμή του a, η εντροπία είναι θετική και μεγιστοποιείται στο a=1/2
  - Εκεί που μεγιστοποιείται και ο βαθμός αβεβαιότητας



Θεωρία Πληροφορίας

- ✓ Ομοίως ορίζεται η απο κοινού αβεβαιότητα:
  - $H(X, Y) = - \Sigma_i \Sigma_j p(x_i, y_j) \log p(x_i, y_j)$
- ✓ Καθώς και η υπο συνθήκη αβεβαιότητα:
  - $H(X|Y) = - \Sigma_i \Sigma_j p(x_i, y_j) \log p(x_i|y_j)$
- ✓ Ιδιότητες:
  - Αν X, Y ανεξάρτητες, τότε $H(X,Y) = H(X) + H(Y)$ (η αβεβαιότητα αθροίζεται)
  - Γενικά, $H(X,Y) \leq H(X) + H(Y)$
  - $H(X|Y) \leq H(X)$ (η αβεβαιότητα μειώνεται όσο μαθαίνουμε)
  - $H(X|Y, Z) \leq H(X|Y)$ (ομοίως)
  - $H(X, Z|Y) \geq H(X|Y)$ (αβεβαιότητα αυξάνεται όταν αυξάνεται το σύνολο των πιθανών ενδεχομένων)
  - **Αμοιβαία πληροφορία:** $I(X, Y) = H(X) - H(X|Y)$ (χρησιμοποιείται στη μέτρηση διαρροής πληροφορίας)

Θεωρία Πληροφορίας

- Μελέτη κρυπτοσυστημάτων με θεωρία πληροφορίας
  - ✓ Έστω ένα τυχαίο κείμενο που θέλουμε να κρυπτογραφήσουμε και έστω ότι επιλέγουμε τυχαία ένα κλειδί με βάση κάποια κατανομή πιθανότητας
  - ✓ P → τυχαία μεταβλητή για την τιμή του plaintext
  - ✓ K → τυχαία μεταβλητή για την τιμή του κλειδιού
  - ✓ C → τυχαία μεταβλητή για την τιμή του ciphertext
  - ✓ Σε όλα τα συμμετρικά κρυπτοσυστήματα:
    - H(C|P, K) = 0 (αν ξέρω το απλό κείμενο και το κλειδί δεν υπάρχει αβεβαιότητα για το ciphertext)
    - H(P|C, K) = 0 (αν ξέρω το κρυπτοκείμενο και το κλειδί, μπορώ να αποκρυπτογραφήσω)

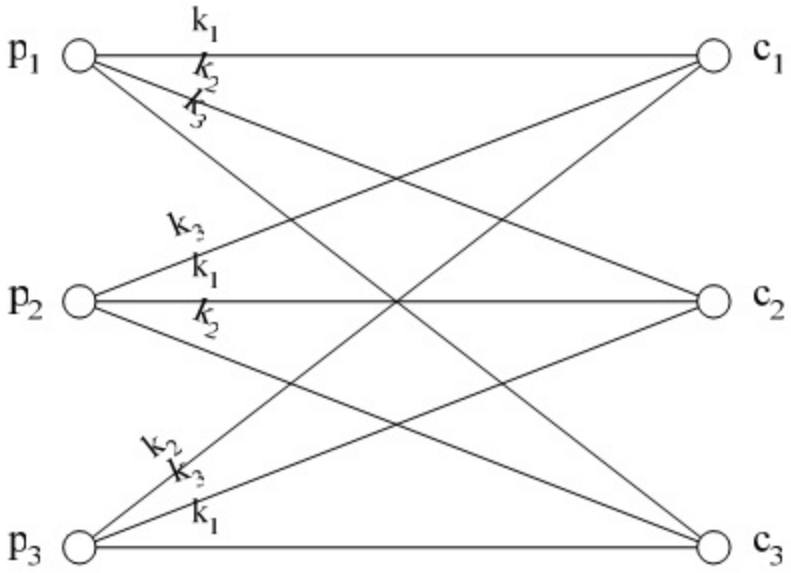
- Τέλεια μυστικότητα (perfect secrecy)

  - ✓ Διαισθητικά: ο Oscar δεν μπορεί να αποκτήσει κάποια πληροφορία για το plaintext (ή για το κλειδί) παρατηρώντας το ciphertext

  - ✓ Ορισμός: Ένα κρυπτοσύστημα είναι *perfectly secure* ή *unconditionally secure against ciphertext-only attacks* αν $H(P|C) = H(P)$

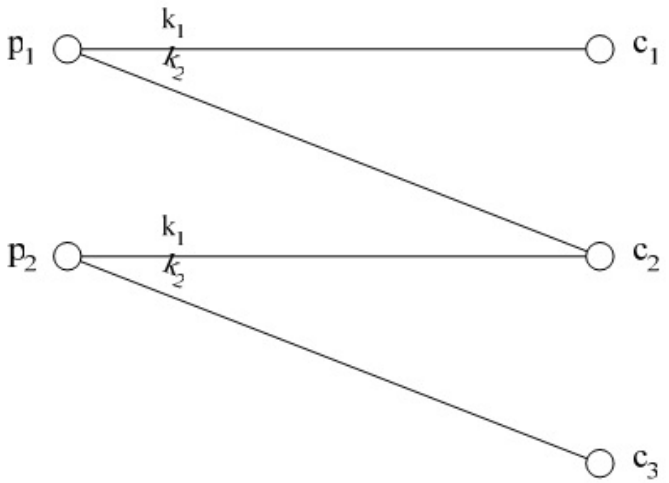
  - ✓ Ισοδύναμα: αν για κάθε πιθανή τιμή x για το P, και κάθε πιθανή τιμή y για το C: $Pr(P = x|C=y) = Pr(P=x)$

- Τέλεια μυστικότητα (perfect secrecy)
  - ✓ Παράδειγμα: έστω ότι έχουμε 3 σύμβολα απλού κειμένου, p_1, p_2 , p_3, που μπορούν να αντιστοιχιστούν σε 3 σύμβολα κρυπτοκειμένου ανάλογα με το κλειδί που διαλέγουμε



Ο Oscar βλέποντας το ciphertext δεν μπορεί να εξάγει κανένα συμπέρασμα για το plaintext

- Τέλεια μυστικότητα (perfect secrecy)
  - ✓ Αντιθέτως, αν το κρυπτοσύστημα ήταν:



- ✓ Όταν ο Oscar βλέπει c_1 ξέρει ότι το plaintext είναι p_1
- ✓ Μόνη αβεβαιότητα όταν βλέπει c_2
- ✓ Άρα δεν είναι perfectly secure

- Τέλεια μυστικότητα (perfect secrecy)

  - ✓ Θεώρημα 1: Έστω ότι στο shift cipher διαλέγουμε ισοπίθανα για κάθε σύμβολο του plaintext ένα τυχαίο κλειδί $k \in Z_{26}$. Τότε για κάθε κατανομή πιθανότητας του plaintext, το κρυπτοσύστημα είναι perfectly secure

  - ✓ Θεώρημα 2: Έστω ότι στο κρυπτοσύστημα one-time pad διαλέγουμε ισοπίθανα για κάθε m-bit string του plaintext ένα τυχαίο κλειδί $k \in \{0,1\}^m$. Τότε για κάθε κατανομή πιθ/τας του plaintext, το κρυπτοσύστημα είναι perfectly secure
    - Κάνοντας XOR με τυχαίο string, πετυχαίνουμε ότι $\Pr[C = y] = (1/2)^m$ για οποιοδήποτε m-bit string y

- Τέλεια μυστικότητα (perfect secrecy)
  - ✓ Έστω K ο χώρος των κλειδιών
  - ✓ Όταν τα plaintext και ciphertext είναι ισομεγέθη (m-bit blocks), για κάθε κλειδί, η συνάρτηση κρυπτογράφησης είναι ένα bijection
  - ✓ Θεώρημα 3: Όταν το plaintext, το ciphertext, και το κλειδί είναι ισομεγέθη και κάθε κλειδί επιλέγεται ισοπίθανα, το σύστημα έχει τέλεια μυστικότητα
    - Ιδέα απόδειξης: κάθε κλειδί προσδιορίζει ένα μοναδικό bijection,
    - Όταν το κλειδί επιλέγεται ισοπίθανα, ουσιαστικά επιλέγεται ένα τυχαίο bijection
    - άρα το ciphertext δεν μπορεί να μειώσει την αβεβαιότητα του plaintext

Τέλεια μυστικότητα (perfect secrecy)

- ✓ Ορισμός: Το *effective key size* του χώρου κλειδιών είναι log $|\boldsymbol{K}|$ (= πόσα bits χρειαζόμαστε για να προσδιορίσουμε ένα κλειδί)
- ✓ Αν τα κλειδιά είναι ισοπίθανα, η *εντροπία* του χώρου κλειδιών, H(K), ισούται με το effective key size
- ✓ Όταν τα plaintext και ciphertext είναι ισομεγέθη (m-bit blocks), τέλεια μυστικότητα έχουμε όταν η εντροπία του χώρου κλειδιών είναι logm
- ✓ Τα συστήματα που παρέχουν τέλεια μυστικότητα είναι πρακτικά ανεφάρμοστα σήμερα
  - Τεράστιο μέγεθος κλειδιού
  - Κόστος παραγωγής, μετάδοσης, αποθήκευσης, επεξεργασίας
- ✓ Στόχος της κρυπτογραφίας: Σχεδιασμός συστημάτων με το καλύτερο trade-off ως προς:
  - Μέγεθος κλειδιού σχετικά μικρό (κρυπτογράφηση μεγάλων ακολουθιών χαρακτήρων χωρίς αλλαγή κλειδιού)
  - Μέγεθος κλειδιού σχετικά μεγάλο (για να μην μπορεί να σπάσει με brute force)
  - Ασφάλεια που να πλησιάζει την τέλεια μυστικότητα

- Πλεονασμός – Περίσσεια Γλώσσας
  - ✓ *Περίσσεια μίας γλώσσας (redundancy):* ποσοστό συνδυασμών γραμμάτων που δεν αντιστοιχούν σε μηνύματα που ανήκουν στη γλώσσα
    - Π.χ. στην ελληνική γλώσσα, με 4 χαρακτήρες, το «χέρι» είναι έγκυρη λέξη αλλά το «ουγκ» δεν είναι
  - ✓ *Απόλυτος ρυθμός γλώσσας* (absolute rate): $A = \lceil \log(n) \rceil$, όπου n το πλήθος των γραμμάτων του αλφάβητου = πόσα bits χρειαζόμαστε για να αναπαραστήσουμε τα γράμματα του αλφαβήτου
  - ✓ Ο αριθμός των δυνατών μηνυμάτων μήκους m, σε γλώσσα με απόλυτο ρυθμό A, είναι 2^{Am}
  - ✓ Έστω 2^{Rm} ο αριθμός των μηνυμάτων μήκους m γραμμάτων που ανήκουν στην γλώσσα (***έγκυρα)***.
  - ✓ Η περίσσεια της γλώσσας ορίζεται ως η ποσότητα: $D=A-R$ (είναι περίσσεια σε bits)
  - ✓ Λατινικό αλφάβητο
    - Το λατινικό αλφάβητο μεταφέρει **log26=4,7 bits/character**, άρα $A = 5$
    - Από μετρήσεις, η μέση ποσότητα πληροφορίας που μεταφέρεται ανά character (per-character entropy) σε κατανοήσιμο Αγγλικό κείμενο είναι 1.5 bits
    - Ο πλεονασμός σε Αγγλικό κείμενο είναι 5−1,5 = 3,5 bits

Πλεονασμός – Περίσσεια Γλώσσας

- Είναι επιθυμητό η περίσσεια *D* μιας γλώσσας να είναι όσο το δυνατόν μικρότερη
- Αν είναι μικρή:
  - Έστω ότι ο αντίπαλος έχει στην κατοχή του το κρυπτοκείμενο
  - κατά την προσπάθεια αποκρυπτογράφησης του Oscar αυτό θα έχει μεγαλύτερη πιθανότητα να αντιστοιχίζεται σε πολλά έγκυρα απλά κείμενα.
- Αν η περίσσεια είναι μεγάλη
  - ο αντίπαλος θα μπορεί πιο εύκολα να αναγνωρίσει το ζητούμενο απλό κείμενο ή να αποκλείσει κλειδιά
  - τα λάθος κλειδιά θα αποκρυπτογραφούν το κρυπτοκείμενο σε μη έγκυρα κείμενα με μεγάλη πιθανότητα.
- Υπάρχει εξάρτηση μεταξύ της περίσσειας της γλώσσας και του μεγέθους του κρυπτοκειμένου που χρειάζεται να έχει ο αντίπαλος για να ανακτήσει το απλό κείμενο

- Πλεονασμός – Περίσσεια Γλώσσας
- [Shannon 1949]:
- **unicity distance (UD)=** η μέση ποσότητα του κρυπτοκειμένου που χρειάζεται για την ανάκτηση του απλού κειμένου
  = το μικρότερο μήκος του κρυπτοκειμένου με το οποίο $H(P/C)=0$.
  - Θεώρημα (Shannon): **UD = H(K) / D**
    - όσο μικρότερη η περίσσεια της γλώσσας, τόσο περισσότερο κρυπτοκείμενο απαιτείται για να εντοπισθεί το κλειδί
  - ✓ Παράδειγμα ενός permutation cipher με αντιμεταθέσεις μήκους t
    - Έστω ότι το plaintext έχει πλεονασμό των $D_{\Pi}=3,2$ bits/character
    - $H(K)/D_{\Pi} = \log(t!)/3,2$
    - Για t=12, χρειαζόμαστε τουλάχιστον 9 χαρακτήρες
    - Για t=27, χρειαζόμαστε τουλάχιστον 29 χαρακτήρες

Ανίχνευση Γλώσσας

- Μέχρι τώρα υποθέσαμε ότι ο Oscar γνωρίζε σε ποια γλώσσα έχει γραφτεί το plaintext
- Αν υποθέσουμε ότι δεν ξέρει τη γλώσσα?
  - ✓ Π.χ. Πολλές λατινογενείς γλώσσες έχουν το ίδιο αλφάβητο
- Έμφυτα χαρακτηριστικά γλώσσας που ένα ασφαλές κρυπτοσύστημα θα πρέπει να έχει την ικανότητα να αποκρύπτει
  - άνιση κατανομή συχνοτήτων των συμβόλων του απλού κειμένου,
  - περίσσεια μιας γλώσσας
- Έλεγχος Κάπα και έλεγχος Χι:
  - εκμεταλλεύονται στατιστικά χαρακτηριστικά κρυπτοκειμένου για να προσδιορίσουν τη γλώσσα γραφής απλού κειμένου

- Έλεγχος Κάπα:
  - ✓ μέτρηση σχετικής συχνότητας εμφάνισης ενός συμβόλου στην ίδια θέση σε διαφορετικά κείμενα
  - ✓ Έστω $M=[m_1 m_2 \ldots m_n]$ και $M'=[m'_1 m'_2 \ldots m'_n]$ δυο διαφορετικά κείμενα κοινού μήκους n.
  - ✓ Η *ποσότητα Κάπα ή ποσότητα ταύτισης* ορίζεται ως:

$\kappa(M, M') = (1/n)\Sigma_i \delta(m_i, m'_i)$ όπου

$$\delta(\alpha, \beta) = \begin{cases} 1 \text{ αν } \alpha=\beta \\ 0 \text{ αν } \alpha\neq\beta \end{cases}$$

  - ✓ Με στατιστικές αναλύσεις έχει βρεθεί ότι η ποσότητα κ εξαρτάται από τη γλώσσα στην οποία είναι τα κείμενα M, M′ και είναι διαφορετική για κάθε γλώσσα

Γλώσσα	Πλήθος γραμμάτων αλφάβητου	Κάπα (%)
Ελληνικά*	24	8,75
Αγγλικά	26	6,61
Γερμανικά	26	7,62
Γαλλικά	26	7,78
Ιταλικά	26	7,38
Ισπανικά	26	7,75
Ιαπωνικά (Romaji)	26	8,19
Ρώσικα	32	5,29

  - ✓ Για μονοαλφαβητικές αντικαταστάσεις και πολυαλφαβητικές γραμμικής φύσης (π.χ. Vigenere), η ποσότητα κ 2 κρυπτοκειμένων (με το ίδιο κλειδί) είναι στατιστικά ίση με την ποσότητα κ των 2 αντίστοιχων plaintexts

- Έλεγχος Χι:
  - ✓ Κρυπταναλυτικά πιο αδύναμος από τον έλεγχο Κάπα
  - ✓ Μέτρηση συχνότητας εμφάνισης συμβόλου σε διαφορετικά κείμενα
  - ✓ Έστω $M=[m_1m_2\ldots m_n]$ και $M'=[m'_1m'_2\ldots m'_n]$ δυο διαφορετικά κείμενα κοινού μήκους n και $\{x_1,x_2,\ldots,x_m\}$ το αλφάβητο των κειμένων
  - ✓ Έστω f_i και f_i' οι συχνότητες εμφάνισης του συμβόλου x_i στα κείμενα M και M′ αντίστοιχα.
  - ✓ Η ποσότητα Χι ορίζεται ως:

$$\chi(M,M')=(1/m^2)\sum_i f_i * f_i'$$

  - ✓ Με στατιστικές αναλύσεις έχει βρεθεί ότι η ποσότητα Χι εξαρτάται από τη γλώσσα στην οποία είναι τα κείμενα M, M′ και είναι διαφορετικό για κάθε γλώσσα
  - ✓ Μόνο για μονοαλφαβητική αντικατάσταση