Πιθανότητες και Αλγόριθμοι

Διδάσκοντες: **Αρ. Παγουρτζής, Δ. Φωτάκης, Δ. Σούλιου, Παν. Γροντάς**

Επιμέλεια διαφανειών: **Δ. Φωτάκης**

Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών

Εθνικό Μετσόβιο Πολυτεχνείο

Πιθανοτικοί Αλγόριθμοι

- **Πιθανοτικός αλγόριθμος** κάνει τυχαίες επιλογές και εξαρτά εξέλιξή του από αυτές.
  - Κατανομή πιθανότητας πάνω σε ντετερμινιστικούς αλγόριθμους.
- Πλεονεκτήματα πιθανοτικών αλγόριθμων:
  - Απλότητα και κομψότητα (π.χ. quickselect, primality).
  - Συνήθως ταχύτεροι από ντετερμινιστικούς.
  - Όταν έχουμε μερική γνώση, περιορισμένη μνήμη, κλπ., πρακτικά αποτελούν μόνη αποδοτική λύση.
- Μειονεκτήματα:
  - Λάθος απάντηση (με μικρή πιθανότητα).
  - Κυμαινόμενος χρόνος εκτέλεσης.
  - Δύσκολο debugging.

Πώς τα Καταφέρνουν;

- Εκμεταλλεύονται «εργαλεία» της πιθανότητας.
- «Αδυνατίζει» (και γίνεται πιο ρεαλιστική) η χειρότερη περίπτωση (π.χ. quicksort).
- Τυχαία δειγματοληψία: αντιπροσωπευτικό δείγμα και λύση (π.χ. clustering, sublinear algs).
- Ικανό πλήθος πιστοποιητικών (βλ. property testing).
- Τυχαία μοιρασιά εργασιών: ισορροπημένη και με ελάχιστο κόστος (υπολογιστικό, επικοινωνιακό).
- Fingerprinting και hashing.
- «Σπάσιμο» συμμετρίας (π.χ. Ethernet, leader election).
- Προσομοίωση διαδικασιών και rapid mixing.

Γινόμενο Πολυωνύμων

- Πολυώνυμα $P_1(x)$ και $P_2(x)$ βαθμού d, και πολυώνυμο $P_3(x)$ βαθμού 2d, όλα ορισμένα σε field F.
- Έλεγχος αν $P_1(x) \times P_2(x) = P_3(x)$
  - … σε χρόνο (σημαντικά) μικρότερο του πολλαπλασιασμού;
- Ελέγχουμε αν $Q(x) = P_1(x) \times P_2(x) - P_3(x)$ είναι (ταυτ.) 0.
  - Έστω Q(x) βαθμού 2d και όχι (ταυτοτικά) 0.
    Τότε, $Pr_{r \in F}[Q(r) = 0] \leq 2d/|F|$.
  - Για $|F| = 200d$ και 3 ανεξ. δείγματα, πιθαν. λάθους $\leq 10^{-6}$.
  - Χρόνος πολ/μού: $\Theta(d^2)$ (ή $\Theta(d \log d)$). Χρόνος ελέγχου: $\Theta(d)$.
- Επεκτείνεται σε πολυώνυμα πολλών μεταβλητών, όπου αντίστοιχη πιθανότητα ορίζεται με συνολικό βαθμό.
  - Θεώρημα Schwartz-Zippel.

Γινόμενο Πινάκων

- Δίνονται A, B, C πίνακες n × n.
  - Έλεγχος αν AB = C σε χρόνο $O(n^2)$.
- Τυχαίο διάνυσμα $r \in \{0, 1\}^n$. Απαντ. NAI ανν A(Br) = Cr.
  - Ισοδύναμα ανν Dr = 0, όπου D = (AB – C).
  - Αν D ≠ 0, D έχει μη μηδενικά στοιχεία.
    Χβτγ., κάποια στην 1η γραμμή του D, ένα στην 1η στήλη.
  - Για κάθε επιλογή των $r_2, \ldots, r_n$,
    υπάρχει μια (το πολύ) επιλογή για το r_1 τ.ω. $\sum_{j=1}^{n} D_{1j} r_j = 0$
  - Άρα πιθανότητα λάθους ≤ 1/2.
  - Με π.χ. 30 ανεξάρτητες επαναλήψεις, πιθ. λάθους < 10^{-6}.

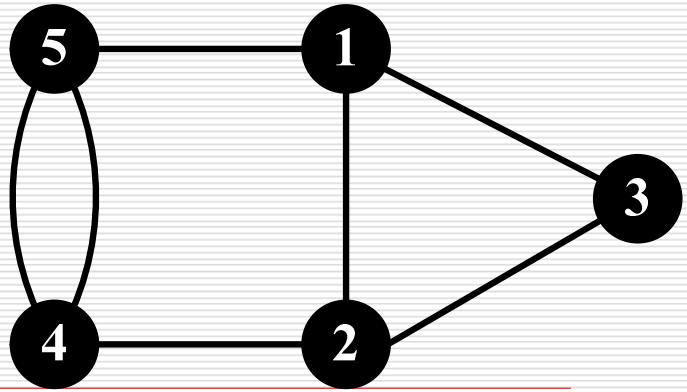
Γινόμενο Πινάκων: Εργαλείο

- Ανάλυση βασίζεται σε αρχή αναβολής τυχαίων αποφάσεων (principle of deferred decisions):
  - «Φιξάρουμε» μέρος των τυχαίων επιλογών (συνήθως σε αυθαίρετες τιμές).
  - Υπολογίζουμε πιθανότητα, δεδομένων αυτών των τιμών.
    - Τεχνικά, υπολογίζουμε την πιθανότητα υπό συνθήκη.
      Επειδή ισχύει για αυθαίρετη συνθήκη, ισχύει χωρίς συνθήκη.
- Γενικότερα, έστω E_1, .., E_n μια διαμέριση του δειγματοχώρου σε γεγονότα. Τότε:

$$\mathbb{P}\mathrm{r}[B] = \sum_{i=1}^{n} \mathbb{P}\mathrm{r}[B \cap E_i] = \sum_{i=1}^{n} \mathbb{P}\mathrm{r}[B|E_i]\,\mathbb{P}\mathrm{r}[E_i]$$

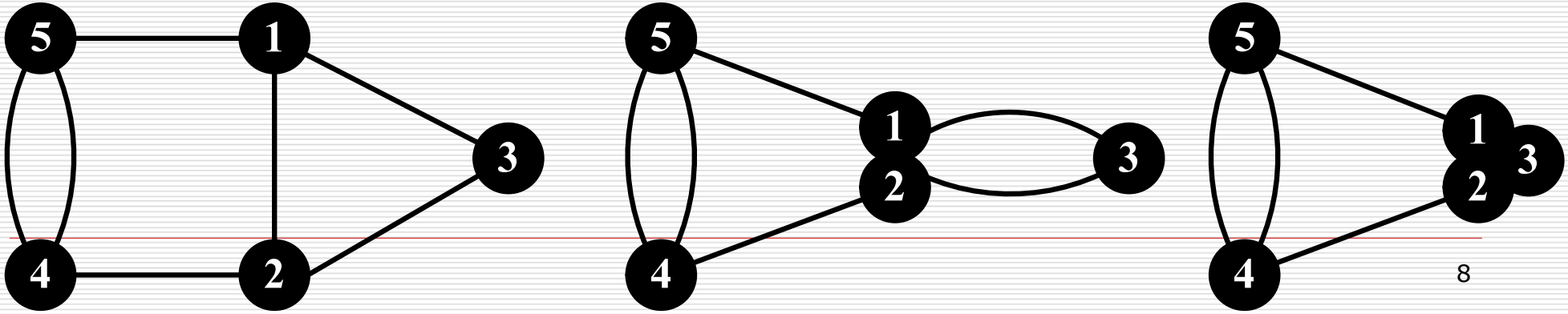
Ελάχιστη Τομή

- Μη κατευθυνόμενο συνεκτικό πολυγράφημα G(V, E).
  - Πολλαπλές ακμές, όχι χωρητικότητες / βάρη.
- Τομή: διαμέριση κορυφών (S, V \ S) με $\emptyset \neq S \subset V$.
  - Σύνολο ακμών που αφαίρεσή τους δημιουργεί τουλ. 2 συνεκτικές συνιστώσες.
  - Μέγεθος τομής $b(S, V \setminus S) = |\{\{u, v\} \in E : u \in S, v \notin S\}|$
- Πρόβλημα: υπολογισμός μιας ελάχιστης τομής.
  - Λύνεται σε χρόνο $O(n^4)$ με διαδοχικές εφαρμογές αλγόριθμου μέγιστης ροής.
  - Υπάρχουν εξειδικευμένοι αλγόριθμοι με χρόνο $O(n^3)$.



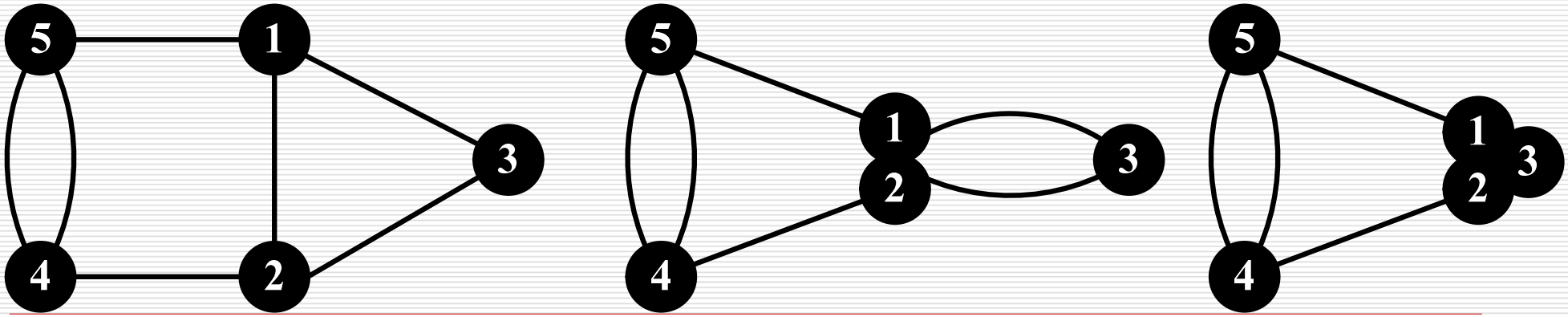
Σύμπτυξη Κορυφών

- **Σύμπτυξη** κορυφών u και v (που συνδέονται με ακμή):
  - Αντικατάσταση u, v από μία νέα κορυφή uv.
  - Κάθε ακμή {x, u} / {x, v} αντικαθίσταται από ακμή {x, uv}.
  - Ακμές {u, v} παραλείπονται.
  - Διαδοχικές συμπτύξεις κορυφών 1, 2 και 12, 3.
- Τομή σε γράφημα μετά από διαδοχικές συμπτύξεις αντιστοιχεί σε τομή σε αρχικό γράφημα.
  - Λειτουργία σύμπτυξης **δεν** μειώνει ελάχιστη τομή.



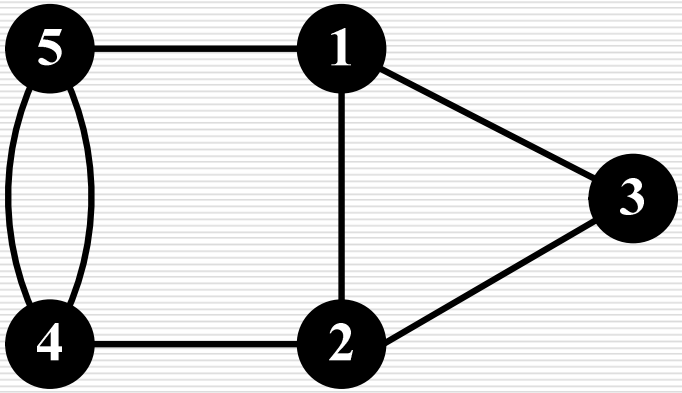
Πιθανοτικός Αλγόριθμος [Karger, 93]

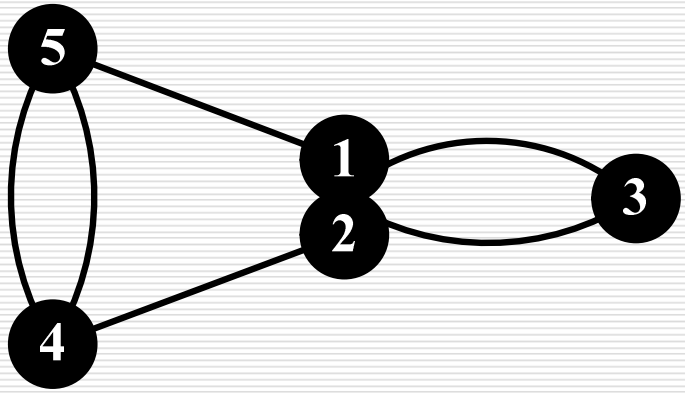
- Ενόσω το γράφημα που απομένει έχει > 2 κορυφές:
  - Διάλεξε μια τυχαία **ακμή** {u, v}.
  - Αντικατέστησε γράφημα με αυτό που προκύπτει από σύμπτυξη κορυφών u και v.
- Ακμές τομής αυτές μεταξύ 2 κορυφών που απομένουν.
- Τομή ορίζεται από κορυφές που συμπτύχθηκαν στις 2 κορυφές που απομένουν.

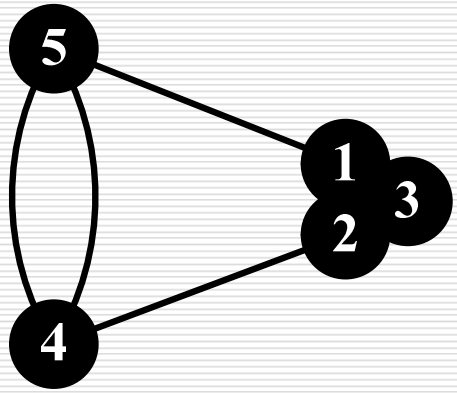


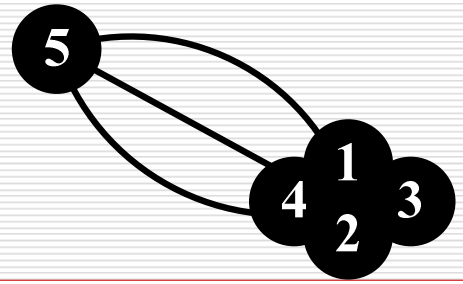
Παράδειγμα

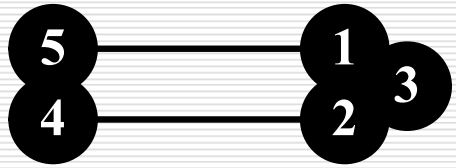
- Αρχικές συμπτύξεις 1, 2, και 12, 3.
  - Σύμπτυξη 123, 4.
  - Σύμπτυξη 5, 4.





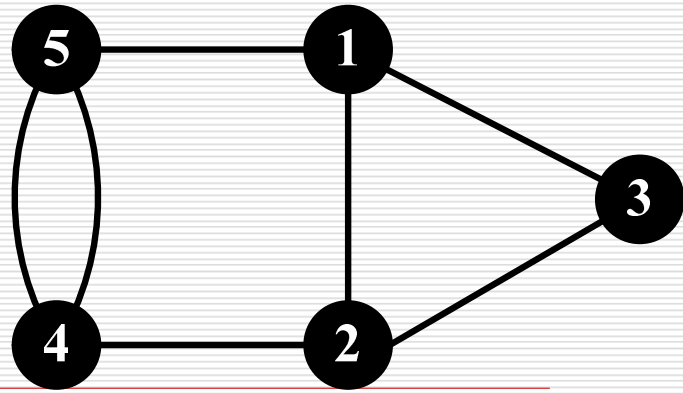






Πιθανοτικός Αλγόριθμος [Karger, 93]

- Βασικές ιδιότητες:
  - Πάντα τερματίζει έπειτα από $n - 2$ συμπτύξεις.
  - Υπολογίζει μία τομή, μπορεί **όχι** ελάχιστη.
  - Ποια πιθανότητα p να καταλήξει σε ελάχιστη τομή;
  - Αν p όχι αμελητέα, μεγαλώνει γρήγορα με επαναλήψεις.
  - Αν $p \geq 2/n^2$, πιθανότητα τουλ. μία από $n^2 \ln n$ επαναλήψεις να καταλήξει σε ελάχιστη τομή $\geq 1 - 1/n^2$.
- Έστω ελάχιστη τομή $C = \{e_1, \ldots, e_k\}$ μεγέθους k.
  - Αλγ. επιστρέφει C ανν καμία από ακμές C δεν επιλεγεί για σύμπτυξη.



Πιθανότητα Επιτυχίας

- Συγκεκριμένη ελάχιστη τομή $C = \{e_1, \ldots, e_k\}$ μεγέθους k.
  - Πιθανότητα καμία από ακμές C **δεν** επιλέγεται για σύμπτυξη.
  - Ελάχιστος βαθμός κορυφής ≥ ελάχιστη τομή.
  - G(V, E) έχει ελάχιστο βαθμό κορυφής ≥ k.
    - G έχει #ακμών ≥ $nk/2$.
    - Πιθανότητα δεν επιλέγεται ακμή του C στην 1η σύμπτυξη:

$$p_1 \geq \frac{\frac{nk}{2} - k}{\frac{nk}{2}} = \frac{n-2}{n}$$

  - Μετά από t συμπτύξεις, γράφημα έχει ελάχιστο βαθμό ≥ k.
    - #ακμών ≥ $(n-t)k/2$.
    - Πιθανότητα δεν επιλέγεται ακμή C του ούτε στην $(t+1)$η σύμπτυξη:

$$p_{t+1} \geq \frac{\frac{(n-t)k}{2} - k}{\frac{(n-t)k}{2}} = \frac{n-t-2}{n-t}$$

Πιθανότητα Επιτυχίας

- Συγκεκριμένη ελάχιστη τομή $C = \{e_1, \ldots, e_k\}$ μεγέθους k.
  - Πιθανότητα καμία από ακμές C δεν επιλέγεται για σύμπτυξη:

$$p = p_1 \cdot p_2 \cdots p_{n-2} \geq \frac{n-2}{n} \cdot \frac{n-3}{n-1} \cdot \frac{n-4}{n-2} \cdots \frac{2}{4} \cdot \frac{1}{3} = \frac{2}{n(n-1)}$$

- Άρα $p \geq 2/n^2$, και πιθανότητα τουλ. μία από $n^2 \log n$ επαναλήψεις να καταλήξει σε **ελάχιστη τομή** $\geq 1 - 1/n^2$.
  - Χρόνος εκτέλεσης $O(n^2)$ / επανάληψη.
  - Συνολικός χρόνος $O(n^4 \log n)$.

Χρόνος Εκτέλεσης

- Όμως (σχετικά) μικρή πιθανότητα αποτυχίας στις πρώτες μισές συμπτύξεις!
  - Π.χ. πιθανότητα να μην συμπτυχθεί καμία ακμή C στις πρώτες (n–3)/2 συμπτύξεις $\geq 1/4$.
  - «Ακριβές» συμπτύξεις είναι «επιτυχημένες».
- Αναδρομική υλοποίηση σε φάσεις:
  - Εκτέλεση βασικού αλγόριθμου για n/2 συμπτύξεις 4 φορές.
  - Συνεχίσουμε αναδρομικά για καθένα από τα αποτελέσματα.
  - Χρόνος εκτέλεσης: $O(n^2 \log n)$ ($O(\log n)$ επίπεδα, $O(n^2)$ / επίπεδο)
  - Έστω $P(n)$ πιθανότητα επιτυχίας για γράφημα n κορυφών:
    - $P(n) = 1 - (1 - P(n/2)/4)^4$, με λύση $P(n) = \Omega(1/\log n)$
- Χρόνος εκτέλεσης συνολικά $O(n^2 \log^3 n)$ για πιθανότητα επιτυχίας $= 1 - O(1/n)$.

Monte Carlo vs Las Vegas

- Monte Carlo αλγόριθμοι (π.χ. min-cut):
  - Μπορεί να δώσουν λάθος απάντηση (με μικρή πιθανότητα), χρόνος εκτέλεσης ντετερμινιστικός (συνήθως!).
  - Πιθανότητα λάθους μπορεί να γίνει πολύ-πολύ μικρή με ανεξάρτητες επαναλήψεις.
  - Προβλήματα απόφασης: one-sided error και two-sided error.
  - Πολυωνυμικοί one-sided error αλγόριθμοι: RP και coRP.
  - Πολυωνυμικοί two-sided error αλγόριθμοι: BPP.
- Las Vegas αλγόριθμοι (π.χ. quicksort, quickselect):
  - Πάντα σωστή απάντηση, χρόνος εκτέλεσης τυχαία μεταβλητή.
  - Πολυωνυμικοί αλγόριθμοι: ZPP.

k-Ικανοποιησιμότητα (k-SAT)

- Λογική πρόταση φ σε k-Συζευκτική Κανονική Μορφή, k-CNF:

  $\varphi \equiv c_1 \wedge \ldots \wedge c_m$, όπου $c_i = \ell_{i_1} \vee \ldots \vee \ell_{i_k}$, με $\ell_{i_j} \in \{x_1, \neg x_1, \ldots, x_n, \neg x_n\}$

  - c_j: όροι. ℓ_{i_j}: literals. #literals σε κάθε όρο $\leq k$.

    Π.χ. για $k = 2$: $(x_1 \vee x_2) \wedge (x_1 \vee \neg x_3) \wedge (\neg x_1 \vee x_2) \wedge (x_2 \vee x_3)$

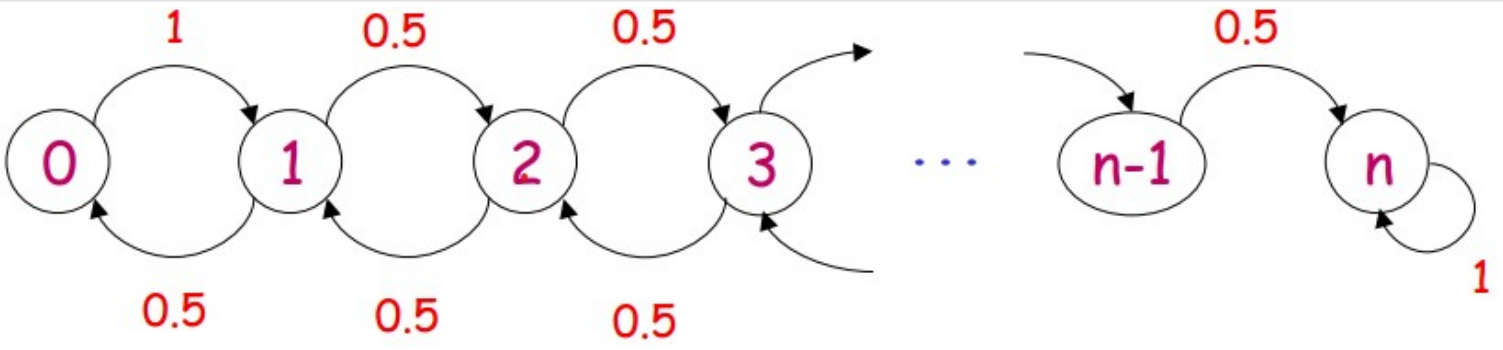
- k-Ικανοποιησιμότητα (k-SAT):
  - Δίνεται φ σε k-CNF. Είναι φ ικανοποιήσιμη;
  - k-SAT είναι NP-πλήρες πρόβλημα για κάθε k ≥ 3.

Πιθανοτικός Αλγόριθμος 2-SAT

- Αρχή με αυθαίρετη ανάθεση τιμών T/F σε μεταβλητές.
- Εκτελούμε για $2qn^2$ φορές και ενόσω φ δεν ικανοποιείται:
  - Έστω όρος *c* που δεν ικανοποιείται.
  - Αλλάζουμε τιμή σε τυχαία μεταβλητή του *c*.
  - Αν *φ* ικανοποιείται, επιστρέφουμε ανάθεση τιμών αλήθειας.
- Αν μετά από $2qn^2$ επαναλήψεις φ δεν ικανοποιείται, συμπεραίνουμε ότι φ μη ικανοποιήσιμη.
- Πιθανότητα λάθος συμπεράσματος, όταν φ ικανοποιήσιμη;
  - Σωστή (τιμή) μετ/της όταν συμφωνεί με τιμή της μετ/της στην ανάθεση που ικανοποιεί *φ*.
  - Αλυσίδα Markov με καταστάσεις #(σωστών μετ/τών).

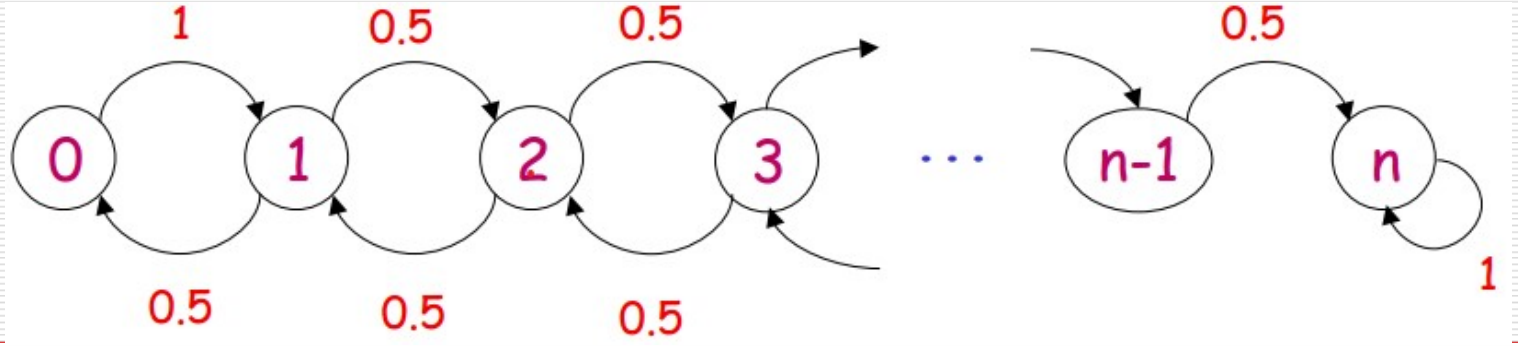
Πιθανοτικός Αλγόριθμος 2-SAT

- Πιθανότητα λάθος συμπεράσματος, όταν φ ικανοποιήσιμη;
  - Markov chain: στοχαστική διαδικασία διακριτού χρόνου όπου πιθανότητες μετάβασης εξαρτώνται μόνο από τρέχουσα κατάσταση (memoryless stochastic process).
  - (Στοχαστικός) πίνακας P με πιθανότητες μετάβασης $P(a, b) = \text{Prob}[X_t = b \mid X_{t-1} = a]$.
  - Αλυσίδα Markov με καταστάσεις #(σωστών μετ/τών).
  - Για κάθε κατάσταση $k = 1, \ldots, n - 1$, $P(k, k+1) = P(k, k-1) = 1/2$.
  - Ασυμμετρία: P(0, 1) = 1, ενώ κατάσταση n τερματική.



Πιθανοτικός Αλγόριθμος 2-SAT

- Πιθανότητα λάθος συμπεράσματος, όταν φ ικανοποιήσιμη;
  - T(k) = μέσος χρόνος για κατάσταση n από τρέχουσα κατάσταση k.
  - T(n) = 0, T(n-1) = 1+T(n-2)/2, …., T(0) = 1+T(1), T(k) = 1 + T(k-1)/2 + T(k+1)/2.
  - Λύση: $T(k) = n^2 - k^2 \leq n^2$
  - Ανισότητα Markov: Prob[όχι κατ. n μετά $2n^2$ βήματα] ≤ ½.
  - Prob[όχι n μετά $2qn^2$ βήματα] ≤ 2^{-q}, Υποδιαίρεση διαδικασίας σε q φάσεις με $2n^2$ βήματα καθεμία.

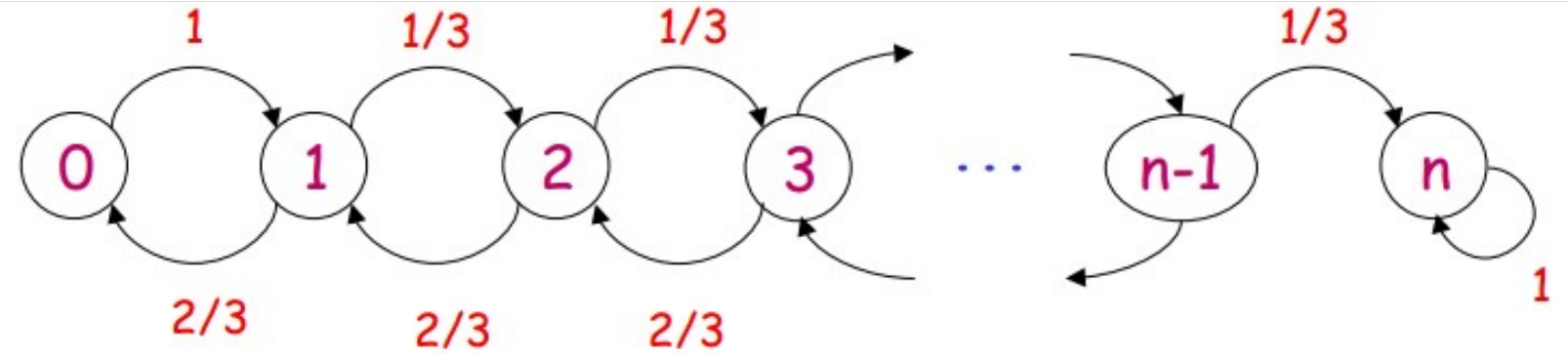


Πιθανοτικός Αλγόριθμος 3-SAT

- Αρχή με **τυχαία** ανάθεση τιμών T/F σε μεταβλητές.
- Εκτελούμε για M φορές και ενόσω φ δεν ικανοποιείται:
  - Έστω όρος *c* που δεν ικανοποιείται.
  - Αλλάζουμε τιμή σε τυχαία μεταβλητή του *c*.
  - Αν *φ* ικανοποιείται, επιστρέφουμε ανάθεση τιμών αλήθειας.
- Αν μετά από M επαναλήψεις φ δεν ικανοποιείται, συμπεραίνουμε ότι φ μη ικανοποιήσιμη.
  - Ενδεχόμενη επανάληψη όλων των παραπάνω για q φορές.
- Πιθανότητα σωστού συμπεράσματος, για φ ικανοποιήσιμη;
  - Αλυσίδα Markov με καταστάσεις #(σωστών μετ/τών).
  - Prob[G] = Prob[τουλάχιστον n/2 σωστές μετ/τές] ≥ ½.

Πιθανοτικός Αλγόριθμος 3-SAT

- Πιθανότητα σωστού συμπεράσματος, για φ ικανοποιήσιμη;
  - Αλυσίδα Markov με καταστάσεις #(σωστών μετ/τών).
  - P(0, 1) = 1, και κατάσταση n τερματική.
  - Για κάθε κατάσταση k = 1, …, n – 1,
    P(k, k+1) = 1/3 και P(k, k – 1) = 2/3.
  - Ισχυρό bias προς 0: χρειάζεται «καλή» αρχική κατάσταση – **αν** πρόκειται να φτάσουμε n, αυτό θα συμβεί σύντομα!
  - Prob[G] Prob[κατ. n | G] ≥ $(1/2)(1/3)^{n/2}$.
  - #επαναλήψεων q για μεγάλη πιθανότητα επιτυχίας:
    $O(3^{n/2} \ln(n)) = O((1.733)^n \ln(n))$.



Πιθανοτικός Αλγόριθμος 3-SAT: Βελτιωμένη Ανάλυση

- Πιθανότητα σωστού συμπεράσματος, για φ ικανοποιήσιμη;
  - Διωνυμική: Prob[αρχικά k σωστές μετ/τές] = $C(n, k)\ 2^{-n}$
  - Πιθανότητα επιτυχίας με k σωστές μετ/τές αρχικά ≥ 3^{-k}
  - Πιθανότητα επιτυχίας συνολικά:

$$\geq \sum_{k=0}^{n} \binom{n}{k} 2^{-n} 3^{-k} = \frac{1}{2^n} \left(1 + \frac{1}{3}\right)^n = \left(\frac{2}{3}\right)^n$$

  - #επαναλήψεων q για μεγάλη πιθανότητα επιτυχίας: $O((1.5)^n \ln(n))$.

Πιθανοτικός Αλγόριθμος 3-SAT: Βελτιωμένη Ανάλυση

- Πιθανότητα σωστού συμπεράσματος, για φ ικανοποιήσιμη;
  - Διωνυμική: Prob[αρχικά k σωστές μετ/τές] = C(n, k) 2^{-n}
  - Θέτουμε M = 3n και υποθέτουμε ότι k σωστές μετ/τές αρχικά.
  - Επιτυχία αν από πρώτα 3k βήματα ≥ 2k «επιτυχημένα».
  - Διωνυμική κατανομή, Prob[2k «σωστά» από 3k]: $\binom{3k}{k}\frac{2^k}{3^{3k}} \geq \frac{1}{2^k\sqrt{6k}}$
  - Πιθανότητα επιτυχίας συνολικά:

$$\geq \sum_{k=1}^{n}\binom{n}{k}2^{-n}\frac{1}{2^k\sqrt{6k}} \geq \frac{1}{2^n\sqrt{6n}}\sum_{k=0}^{n}\binom{n}{k}2^{-k}$$

$$= \frac{1}{2^n\sqrt{6n}}\left(1+\frac{1}{2}\right)^n = \frac{1}{\sqrt{6n}}\left(\frac{3}{4}\right)^n$$

  - #επαναλήψεων q για μεγάλη πιθανότητα επιτυχίας: $O((4/3)^n \ln(n))$.

Μπάλες και Κουτιά

- Έχουμε m μπάλες και n κουτιά. Κάθε μπάλα επιλέγει το κουτί της ισοπίθανα και ανεξάρτητα.
  - Απλό μοντέλο, πλήθος εφαρμογών(!).
  - Μέγιστος #μπαλών σε κάποιο κουτί;
    - Load balancing. Hashing with chains.
  - Ελάχιστο m ώστε να εμφανιστεί κουτί με ≥ 2 μπάλες;
    - Birthday paradox.
  - Ελάχιστο m ώστε κανένα κουτί άδειο;
    - Coupon collecting.

Μέγιστος #Μπαλών

- Πιθανότητα να βρεθεί κουτί με $\geq 3 \ln n / \ln\ln n$ μπάλες είναι $\leq 1/n$.
  - L_i = #μπαλών σε κουτί i: $\mathbb{P}\mathrm{r}[L_i \geq k] \leq \binom{n}{k}\left(\frac{1}{n}\right)^k$ $\qquad k! \geq (k/e)^k$

$$\leq \frac{n^k e^k}{k^k n^k} = \left(\frac{e}{k}\right)^k$$

  - Συνεπώς $\mathbb{P}\mathrm{r}\left[L_i \geq \frac{3\ln n}{\ln\ln n}\right] \leq n^{-2}$
  - ... και (από union bound) $\mathbb{P}\mathrm{r}\left[\exists i : L_i \geq \frac{3\ln n}{\ln\ln n}\right] \leq \frac{n}{n^2} = \frac{1}{n}$
  - Πιο ακριβής ανάλυση είναι εφικτή [Gonnet].
- Ndo με πιθανότητα $\geq 1 - 1/n$, υπάρχει κουτί με $\Omega(\ln n / \ln\ln n)$ μπάλες.

Συλλογή Κουπονιών

- Ελάχιστο m ώστε κανένα κουτί άδειο.
  - Z_k = #μπαλών όταν για πρώτη φορά #γεμάτων κουτιών = k.
  - $X_k = Z_{k+1} - Z_k$: #μπαλών για να γεμίσει το k+1 κουτί.
  - X_k ακολουθεί γεωμετρική κατανομή με παράμετρο 1 – k/n, και έχει $E[X_k] = n/(n - k)$.
  - Γραμμικότητα μέσης τιμής: $\mathbb{E}[Z_n] = \sum_{k=0}^{n-1} \mathbb{E}[X_k] = \sum_{k=0}^{n-1} \frac{n}{n-k} = nH_n$
- Εμφανίζει ισχυρή συγκέντρωση γύρω από την μέση τιμή:
  - $Y_{j,k}$: κουτί j είναι άδειο μετά τις πρώτες k μπάλες.
  - Για κάθε β > 1, πιθανότητα κάποιο κουτί άδειο μετά από $\beta n \ln n$ μπάλες:
  - Μπορεί ν.δ.ο. για κάθε c, πιθανότητα κάποιο κουτί άδειο μετά από $n(\ln n + c)$ μπάλες:

$$\mathbb{Pr}[Y_{j,k}] = \left(1 - \frac{1}{n}\right)^k \leq e^{-k/n}$$

$$\leq n\, e^{-\beta \ln n/n} = n^{1-\beta}$$

$$\leq e^{-e^{-c}}$$

Πιθανοτική Μέθοδος

- Μέθοδος κατασκευής υπαρξιακών αποδείξεων.
  - Γενικεύει αρχή περιστερώνα, με χρήση πιθανοτήτων.
- Θέλουμε νδο «υπάρχει αντικείμενο με ιδιότητα Χ».
  - Ορίζουμε δειγματοχώρο που περιέχει αντικείμενα με ιδιότητα Χ (και συνήθως πολλά άλλα αντικείμενα).
  - Αποδεικνύουμε ότι υπάρχει θετική πιθανότητα να επιλέξουμε αντικείμενο με ιδιότητα Χ.
- Ειδική περίπτωση μεθόδου: υπάρχει λύση με αντικειμενική τιμή τουλάχιστον (το πολύ) Χ.
  - Αποδεικνύουμε ότι η μέση τιμή (κατάλληλα) επιλεγμένης κατανομής εφικτών λύσεων είναι τουλάχιστον (το πολύ) Χ.

Διμερή Υπογραφήματα

- Κάθε γράφημα G(V, E) με m ακμές περιέχει διμερές υπογράφημα G'(X, Y, E') με τουλάχιστον m/2 ακμές.
  - Βλ. και πρόβλημα MAX CUT. Ισχύει και για πολυγραφήματα χωρίς ανακυκλώσεις.
- Απόδειξη με πιθανοτική μέθοδο:
  - Κάθε κορυφή στο X με πιθανότητα 1/2, διαφορετικά στο Y.
  - ∀ακμή {u, v}, Prob[{u, v} μεταξύ X και Y] = 1/2.
  - Γραμμικότητα μέσης τιμής: Exp[#ακμών μεταξύ X και Y] = m/2
  - Άρα υπάρχει διαμέριση (X, Y) ώστε #ακμών μεταξύ X και Y ≥ m/2
- Κατασκευαστική απόδειξη (conditional expectations):
  - Εξετάζουμε κορυφές μία-μία με τη σειρά. Κορυφή u στο X αν έχει πιο πολλούς γείτονες στο Y από ότι στο X, διαφορετικά στο Y.
  - «Κρατάμε» μεταξύ X και Y τουλάχιστον τόσες ακμές όσες «διώχνουμε». Τυπική απόδειξη με επαγωγή στον #κορυφών.

Μεγάλα Ανεξάρτητα Σύνολα

- Κάθε γράφημα G(V, E) με n κορυφές και m = nd/2 ακμές, για κάποιο d ≥ 1, έχει ανεξάρτητο σύνολο μεγέθους ≥ n/(2d)
- Απόδειξη με πιθανοτική μέθοδο:
  - Τυχαίο υποσύνολο $V_1 \subseteq V$: κάθε κορυφή u στο V_1 με πιθανότητα p.
  - $Exp[|V_1|] = np$ και $Exp[|E(G[V_1])|] = mp^2 = ndp^2/2$.
  - $G[V_1]$ (πιθανότατα) **δεν** είναι ανεξάρτητο σύνολο.
  - Αλλά αφαιρώντας αναμενόμενο #κορυφών ≤ $ndp^2/2$ από V_1 (το ένα άκρο κάθε ακμής στο $G[V_1]$) παίρνουμε ανεξάρτητο σύνολο.
  - Exp[#κορυφών που μένουν] ≥ $np - ndp^2/2 = np(1 - dp/2)$.
  - Αυτό μεγιστοποιείται για p = 1/d, και έχουμε Exp[#κορυφών που μένουν] ≥ n/(2d)
  - Άρα υπάρχει ανεξάρτητο σύνολο με ≥ n/(2d) κορυφές.
- Κάθε γράφημα G(V, E) έχει ανεξάρτητο σύνολο μεγέθους $\geq \sum_{v \in V} \frac{1}{\deg(v) + 1}$

Συγκέντρωση στη Μέση Τιμή

- (Πραγματική τιμή) «ομαλών» συναρτήσεων μεγάλου αριθμού ανεξάρτητων τυχαίων μεταβλητών «κινείται» σε ένα μικρό διάστημα γύρω από την μέση τιμή.
  - Βλ. [Dubhashi and Panconessi, Concentration of Measure for the Analysis of Randomized Algorithms, 2007].
- Ανισότητα Markov (γενική, αλλά όχι ιδιαίτερα ισχυρή):
  - X μη-αρνητική τυχαία μεταβλητή. Για κάθε t > 0, $\mathbb{P}\mathrm{r}[X \geq t\,\mathbb{E}[X]] \leq 1/t$ $\mathbb{P}\mathrm{r}[X \geq t] \leq \mathbb{E}[X]/t$
- Ανισότητα Chebyshev (γενική, ισχυρότερη):
  - Για κάθε t > 0, $\mathbb{P}\mathrm{r}[|X - \mathbb{E}[X]| \geq t\,\sigma_X] \leq 1/t^2$
  - Απόδειξη εύκολα από ορισμό Var[X] και ανισότητα Markov.

Chernoff Bounds

$$\forall \varepsilon \in (0, 0.7),\ \frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}} \leq 1 - \frac{\varepsilon^2}{e}$$

- Έστω $X_1, \ldots, X_n$ **ανεξάρτητες** Bernoulli τ.μ. με $E[X_k] = p_k$, $X = X_1 + \ldots + X_n$, και $E[X] = \mu$. Για κάθε $\varepsilon > 0$,

$$\mathbb{P}\mathrm{r}[X > (1+\varepsilon)\mu] \leq \left[\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right]^{\mu}$$

  - Για κάθε $t > 0$, και χρησιμοποιώντας ανισότητα Markov:

$$\mathbb{P}\mathrm{r}[X > (1+\varepsilon)\mu] = \mathbb{P}\mathrm{r}[e^{tX} > e^{t(1+\varepsilon)\mu}] \leq \mathbb{E}[e^{tX}]/e^{t(1+\varepsilon)\mu}$$

$$\mathbb{E}[e^{tX}] = \prod_{k=1}^{n} \mathbb{E}[e^{tX_k}] \leq \prod_{k=1}^{n} e^{p_k(e^t-1)} = e^{(e^t-1)\mu}$$

$$\mathbb{P}\mathrm{r}[X > (1+\varepsilon)\mu] \leq \left[\frac{e^{e^t-1}}{e^{t(1+\varepsilon)}}\right]^{\mu}$$

$$\overset{t=\ln(1+\varepsilon)}{\Rightarrow} \mathbb{P}\mathrm{r}[X > (1+\varepsilon)\mu] \leq \left[\frac{e^{\varepsilon}}{(1+\varepsilon)^{(1+\varepsilon)}}\right]^{\mu}$$

Chernoff Bounds

- Έστω $X_1, \ldots, X_n$ **ανεξάρτητες** Bernoulli τ.μ., $X = X_1+\ldots+X_n$, και $E[X] = \mu$.
  - Για κάθε $1 \geq \varepsilon \geq 0$, $\mathbb{P}\mathrm{r}[X > (1+\varepsilon)\mu] \leq e^{-\varepsilon^2\mu/3}$
    $\mathbb{P}\mathrm{r}[X < (1-\varepsilon)\mu] \leq e^{-\varepsilon^2\mu/2}$
  - Εξαιρετικά ισχυρή συγκέντρωση γύρω από την μέση τιμή!
  - Απαιτούν σύγκριση X με λογαριθμική ποσότητα για να «δουλέψουν καλά».
  - Αντίστοιχα φράγματα για τ.μ. X_k με πεδίο τιμών το $[0, w_k]$.
  - Απαιτούν **ανεξαρτησία** (ή αρνητική εξάρτηση).
  - Αντίστοιχα bounds για τ.μ. με περιορισμένη εξάρτηση.
  - Πολύ σημαντικά για την ανάλυση πιθανοτικών αλγόριθμων.

Παραδείγματα

- Αν μοιράζουμε $m = n \ln n$ μπάλες σε n κουτιά, πιθανότητα προκύψει κουτί με $> 3 \ln n$ μπάλες είναι $\leq 1/n$.
- Set Balancing:
  - $A_1, \ldots, A_n$ υποσύνολα U, $|U| = n$ και για κάθε j, $|A_j| = n/2$.
  - Ζητείται διαμέριση U σε B και W που ελαχιστοποιεί: $\max_j ||A_j \cap B| - |A_j \cap W||$
  - Τυχαία διαμέριση B, W: $\max_j ||A_j \cap B| - |A_j \cap W|| \leq 3\sqrt{n \ln n}$ με πιθανότητα $\geq 1 - 2/n^2$.
  - Για κάθε j, $X_j = |A_j \cap W|$ με $E[X_j] = n/4$. Έχουμε:

$$\mathbb{P}\mathrm{r}\left[||A_j \cap B| - |A_j \cap W|| > 3\sqrt{n \ln n}\right] = \mathbb{P}\mathrm{r}\left[|n/2 - 2|A_j \cap W|| > 3\sqrt{n \ln n}\right]$$

$$= \mathbb{P}\mathrm{r}\left[|\mathbb{E}[X_j] - X_j| > \tfrac{3}{2}\sqrt{n \ln n}\right]$$

$$\leq 2\, e^{-\frac{n}{12}(\frac{6\sqrt{\ln n}}{\sqrt{n}})^2} = 2/n^3$$

$$\overbrace{\frac{n}{4}}^{\mathbb{E}[X_j]} \cdot \overbrace{\frac{6\sqrt{\ln n}}{\sqrt{n}}}^{\varepsilon} = \frac{3}{2}\sqrt{n \ln n}$$

Τυχαία Δειγματοληψία

- Σύνολο A, $|A| = n$, (άγνωστου μεγέθους) σύνολο $X \subseteq A$ με στοιχεία A που έχουν κάποια ιδιότητα.
  - Έστω $|X| = p\,n$. Θα υπολογίσουμε εκτίμηση p' για p.
  - Επιλέγουμε «δείγμα» A', $|A'| \geq 3\ln(2/\delta)/\varepsilon^2$, και υπολογίζουμε $p' = |A' \cap X| / |A'|$.
  - Με πιθανότητα $\geq 1 - \delta$, εκτίμηση $p' \in [p-\varepsilon, p+\varepsilon]$.
- Σύνολο A, $|A| = n$, με διαμέριση $A_1, \ldots, A_k$, $|A_j| = a_j n \geq \gamma n$, που ορίζεται από κάποιες ιδιότητες (π.χ. τι ψηφίζουν).
  - Θα εκτιμήσουμε όλα τα a_j γνωρίζοντας μόνο ότι είναι $\geq \gamma$.
  - Επιλέγουμε «δείγμα» B, $|B| \geq 3\ln(2/(\delta\gamma))/(\gamma\varepsilon^2)$.
  - Έστω $B_j = A_j \cap B$ και $\beta_j = |B_j|/|B|$.
  - Με πιθανότητα $\geq 1 - \delta$, για όλα τα A_j, $(1-\varepsilon)a_j \leq \beta_j \leq (1+\varepsilon)a_j$