

# Πιθανότητες και Αλγόριθμοι

---

**Δημήτρης Φωτάκης**

Σχολή Ηλεκτρολόγων Μηχανικών  
και Μηχανικών Υπολογιστών

Εθνικό Μετσόβιο Πολυτεχνείο



# Πιθανοτικοί Αλγόριθμοι

---

- **Πιθανοτικός αλγόριθμος** κάνει **τυχαίες επιλογές** και εξαρτά **εξέλιξη 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

---

- Πολυώνυμο  $P_1(x)$  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 πολυώνυμο **πολλών μεταβλητών**, όπου αντίστοιχη πιθανότητα ορίζεται με **συνολικό βαθμό**.
  - Θεώρημα **Schwartz-Zippel**.

# Γινόμενο Πινάκων

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

# Γινόμενο Πινάκων: Εργαλείο

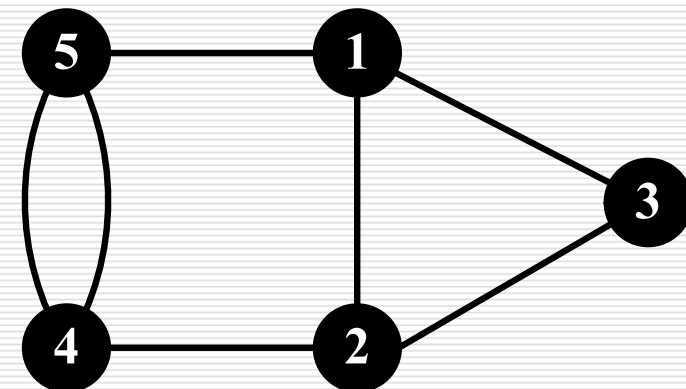
---

- 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  $E_1, \dots, E_n$  μια **διαμέριση** του δειγματοχώρου σε γεγονότα. Τότε:

$$\Pr[B] = \sum_{i=1}^n \Pr[B \cap E_i] = \sum_{i=1}^n \Pr[B|E_i] \Pr[E_i]$$

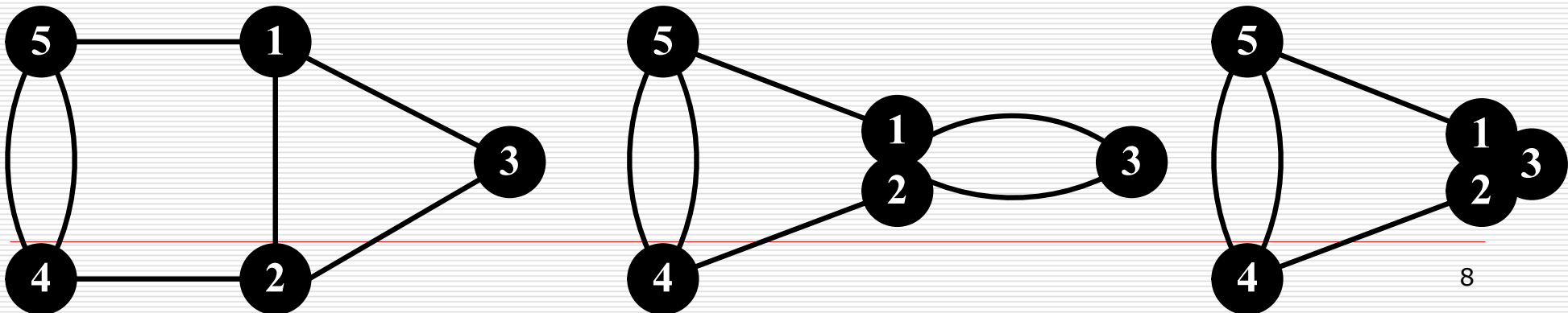
# Ελάχιστη Τομή

- Μη κατευθυνόμενο συνεκτικό **πολυγράφημα**  $G(V, E)$ .
  - Πολλαπλές ακμές, όχι χωρητικότητες / βάρη.
- **Τομή**: διαμέριση κορυφών  $(S, V \setminus S)$  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

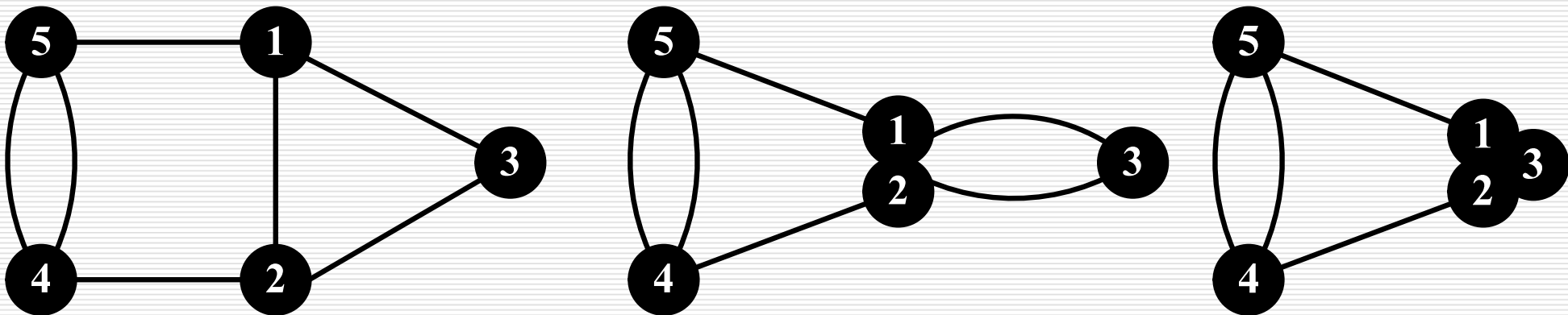
- **Σύμπτυξη** κορυφών **u και v** (που συνδέονται με ακμή):
  - Αντικατάσταση u, v από μία **νέα κορυφή uv**.
  - Κάθε ακμή  $\{x, u\} / \{x, v\}$  αντικαθίσταται από ακμή  $\{x, uv\}$ .
  - Ακμές  $\{u, v\}$  παραλείπονται.
  - Διαδοχικές συμπτύξεις κορυφών 1, 2 και 12, 3.
- **Τομή** σε γράφημα **μετά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





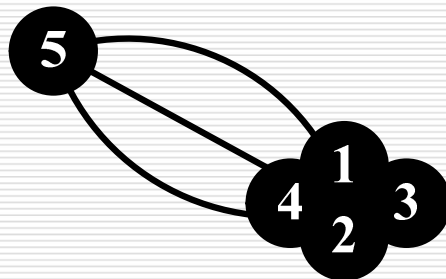
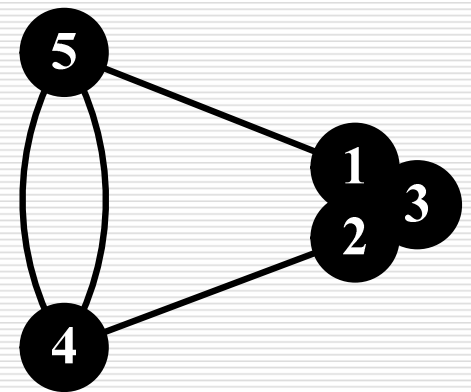
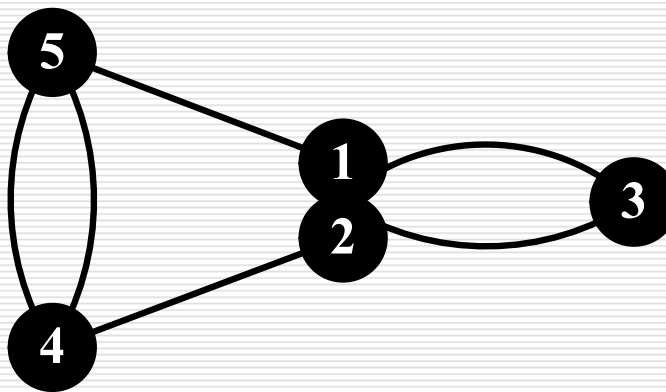
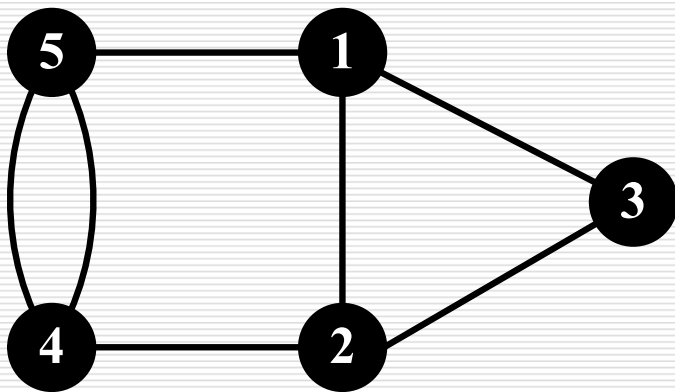
# Πιθανοτικός Αλγόριθμος [Karger, 93]

- **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.



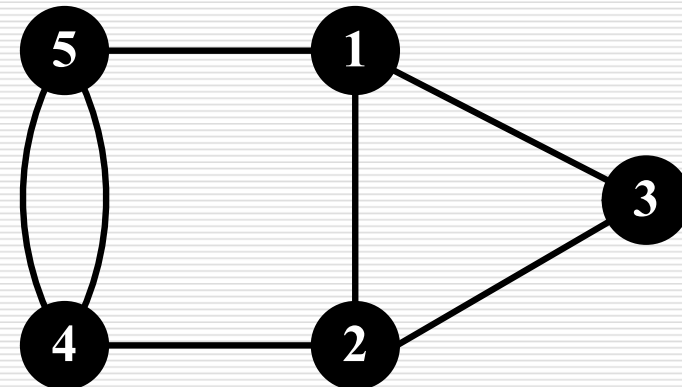
# Παράδειγμα

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



# Πιθανοτικός Αλγόριθμος [Karger, 93]

- Βασικές ιδιότητες:
  - Πάντα **τερματίζει** έπειτα από  $n - 2$  συμπτώξεις.
  - 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  $C = \{e_1, \dots, e_k\}$  μεγέθους  $k$ .
  - Αλγ. επιστρέφει  $C$  ανν καμία από ακμές  $C$  δεν επιλεγεί για σύμπτυξη.



# Πιθανότητα Επιτυχίας

- Συγκεκριμένη ελάχιστη τομή  $C = \{e_1, \dots, e_k\}$  μεγέθους  $k$ .
  - Πιθανότητα **καμία** από ακμές  $C$  **δεν** επιλέγεται για σύμπτυξη.
  - Ελάχιστος βαθμός κορυφής  $\geq$  ελάχιστη τομή.
  - $G(V, E)$  έχει **ελάχιστο βαθμό** κορυφής  $\geq k$ .
    - $G$  έχει **#ακμών**  $\geq nk/2$ .
    - Πιθανότητα **δεν** επιλέγεται ακμή του  $C$  στην **1<sup>η</sup>** σύμπτυξη:
$$p_1 \geq \frac{\frac{nk}{2} - k}{\frac{nk}{2}} = \frac{n-2}{n}$$
    - Μετά από  $t$  συμπτώξεις, γράφημα έχει **ελάχιστο βαθμό**  $\geq k$ .
      - **#ακμών**  $\geq (n-t)k/2$ .
      - Πιθανότητα **δεν** επιλέγεται ακμή  $C$  του **ούτε** στην **( $t+1$ )<sup>η</sup>** 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  $C = \{e_1, \dots, e_k\}$  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 συμπτώξεις!
  - Π.χ. πιθανότητα να μην συμπτυχθεί καμία ακμή  $C$  στις πρώτες  $(n-3)/2$  συμπτώξεις  $\geq 1/4$ .
  - «Ακριβές» συμπτώξεις είναι «επιτυχημένες».
- Αναδρομική υλοποίηση σε φάσεις:
  - Εκτέλεση βασικού αλγόριθμου για  $n/2$  συμπτώξεις 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**.

# Βασικές Έννοιες Πιθανότητας

---

- Δειγματοχώρος, γεγονός και πιθανότητα, τυχαία μεταβλητή.
  - $\Pr[A \cup B] = \Pr[A] + \Pr[B] - \Pr[A \cap B]$   
(γενικεύεται με μέθοδο εγκλεισμού – αποκλεισμού).
  - Union bound:  $\Pr[\cup_{i=1}^n A_i] \leq \sum_{i=1}^n \Pr[A_i]$
  - Πιθανότητα A υπό συνθήκη B:  $\Pr[A|B] = \Pr[A \cap B] / \Pr[B]$   
Γενίκευση:  $\Pr[\cap_{i=1}^n A_i] = \Pr[A_1] \Pr[A_2|A_1] \cdots \Pr[A_n | \cap_{i=1}^{n-1} A_i]$
  - Ανεξάρτητα γεγονότα:  $\Pr[A \cap B] = \Pr[A] \Pr[B]$ .
  - Αρνητικά σχετιζόμενα γεγονότα.



# Βασικές Έννοιες Πιθανότητας

- Μέση τιμή:  $\mathbb{E}[X] = \sum_{k=0}^{\infty} \Pr[X = k] k$ 
  - Ισοδύναμα (ακέραιες τυχαίες μεταβλ.):  $\mathbb{E}[X] = \sum_{k=1}^{\infty} \Pr[X \geq k]$
  - Γραμμικότητα:  $E[X+Y] = E[X] + E[Y]$ .
  - Ανισότητα Jensen: Αν  $f$  κυρτή συνάρτηση,  $E[f(X)] \geq f(E[X])$ .
  - Αν  $X$  και  $Y$  ανεξάρτητες:  $E[X Y] = E[X] E[Y]$ .
- Διακύμανση (variance):
  - $\text{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$
  - Τυπική απόκλιση (std deviation):  $\sigma_x = \text{Var}(X)^{1/2}$
  - Αν  $X$  και  $Y$  ανεξάρτητες:  $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$ .
- Probability generating function:
$$G_X(z) = \sum_{k=0}^{\infty} \Pr[X = k] z^k$$
$$\mathbb{E}[X] = G'(1)$$
$$\text{Var}[X] = G''(1) + G'(1) - G'(1)^2$$

# Παραδείγματα Κατανομών

---

- Bernoulli μεταβλητή  $X$ : 1 με **πιθ.  $p$** , και 0 διαφ.
  - $E[X] = p$ ,  $\text{Var}[X] = p(1 - p)$ ,  $G_X(z) = 1 - p + pz$ .
- Δυωνυμική κατανομή  $\text{Pr}[X = k] = \binom{n}{k} p^k (1 - p)^{n-k}$   
**Bin( $n, p$ ):**
  - **Αριθμός επιτυχιών** σε  $n$  «ρίψεις» με πιθανότητα επιτυχίας  $p$ .
  - Άθροισμα  $n$  Bernoulli μεταβλητών με παράμετρο  $p$ .
  - $E[X] = np$ ,  $\text{Var}[X] = np(1 - p)$ ,  $G_X(z) = (1 - p + pz)^n$
- Γεωμετρική κατανομή **Geo( $p$ ):**  $\text{Pr}[X = k] = (1 - p)^{k-1} p$ 
  - **Αριθμός «ρίψεων»** μέχρι την πρώτη **επιτυχία** (waiting time).
  - $E[X] = 1/p$ ,  $\text{Var}[X] = (1 - p)/p^2$ ,  $G_X(z) = pz / (1 - z + pz)$
  - **Αμνησία:**  $\text{Pr}[X = n+k \mid X > k] = \text{Pr}[X = n]$

# 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  $\geq 2$  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  $i$ :  $\Pr[L_i \geq k] \leq \binom{n}{k} \left(\frac{1}{n}\right)^k \quad k! \geq (k/e)^k$

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

■ Συνεπώς  $\Pr\left[L_i \geq \frac{3 \ln n}{\ln \ln n}\right] \leq n^{-2}$

■ ... και (από union bound)  $\Pr\left[\exists i : L_i \geq \frac{3 \ln n}{\ln \ln n}\right] \leq \frac{n}{n^2} = \frac{1}{n}$

■ Πιο ακριβής ανάλυση είναι εφικτή [Gonnet].

□ Νδο με πιθανότητα  $\geq 1 - 1/n$ , υπάρχει κουτί με  $\Omega(\ln n / \ln \ln n)$  μπάλες.

# Δυο Μπάλες στο Ίδιο Κουτί

---

- Πιθανότητα όλες οι  $m$  ( $< n$ ) μπάλες σε διαφορετικό κουτί:

$$P_m = \frac{n}{n} \frac{n-1}{n} \frac{n-2}{n} \dots \frac{n-m+1}{n} = \prod_{k=1}^{m-1} \left(1 - \frac{k}{n}\right)$$
$$\leq \prod_{k=1}^{m-1} e^{-k/n} = e^{-m(m-1)/(2n)}$$

- Πιθανότητα τουλ. 2 μπάλες στο ίδιο κουτί  $\geq 1 - P_m$ 
  - Για  $n = 365$  και  $m = 28$ : πιθανότητα σε 28 ανθρώπους, κάποιοι να έχουν γενέθλια την ίδια μέρα  $> 1 - e^{-1}$

# 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:  $E[Z_n] = \sum_{k=0}^{n-1} E[X_k] = \sum_{k=0}^{n-1} \frac{n}{n - k} = nH_n$
- Εμφανίζει **ισχυρή συγκέντρωση** γύρω από την μέση τιμή:
  - $Y_{j,k}$ : κουτί  $j$  είναι άδειο μετά τις πρώτες  $k$  μπάλες.  $\Pr[Y_{j,k}] = \left(1 - \frac{1}{n}\right)^k \leq e^{-k/n}$
  - Για κάθε  $\beta > 1$ , πιθανότητα κάποιο κουτί άδειο μετά από  $\beta n \ln n$  μπάλες:  $\leq n e^{-\beta \ln n / n} = n^{1-\beta}$
  - Μπορεί ν.δ.ο. για κάθε  $c$ , πιθανότητα κάποιο κουτί άδειο μετά από  $n(\ln n + c)$  μπάλες:  $\leq e^{-e^{-c}}$

# Πιθανοτική Μέθοδος

---

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

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

# 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.  $\Pr[X \geq t \mathbb{E}[X]] \leq 1/t$   
Για κάθε  $t > 0$ ,  $\Pr[X \geq t] \leq \mathbb{E}[X]/t$
- **Ανισότητα Chebyshev** (γενική, ισχυρότερη):
  - Για κάθε  $t > 0$ ,  $\Pr[|X - \mathbb{E}[X]| \geq t \sigma_X] \leq 1/t^2$
  - Απόδειξη εύκολα από ορισμό  $\text{Var}[X]$  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  $X_1, \dots, X_n$  **ανεξάρτητες** Bernoulli τ.μ. με  $E[X_k] = p_k$ ,  $X = X_1 + \dots + X_n$ , και  $E[X] = \mu$ . Για κάθε  $\varepsilon > 0$ ,

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

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

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

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

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

$$\stackrel{t=\ln(1+\varepsilon)}{\Rightarrow} \Pr[X > (1 + \varepsilon)\mu] \leq \left[ \frac{e^\varepsilon}{(1 + \varepsilon)^{(1+\varepsilon)}} \right]^\mu$$

# Chernoff Bounds

---

- Έστω  $X_1, \dots, X_n$  **ανεξάρτητες** Bernoulli τ.μ.,  
 $X = X_1 + \dots + X_n$ , και  $E[X] = \mu$ .
  - Για κάθε  $1 \geq \varepsilon \geq 0$ ,  
$$\Pr[X > (1 + \varepsilon)\mu] \leq e^{-\varepsilon^2 \mu / 3}$$
$$\Pr[X < (1 - \varepsilon)\mu] \leq e^{-\varepsilon^2 \mu / 2}$$
  - 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:
  - $A_1, \dots, A_n$  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:

$$\begin{aligned} \Pr \left[ ||A_j \cap B| - |A_j \cap W|| > 3\sqrt{n \ln n} \right] &= \Pr \left[ |n/2 - 2|A_j \cap W|| > 3\sqrt{n \ln n} \right] \\ &= \Pr \left[ |\mathbb{E}[X_j] - X_j| > \frac{3}{2}\sqrt{n \ln n} \right] \\ &\leq 2 e^{-\frac{n}{12} \left( \frac{6\sqrt{\ln n}}{\sqrt{n}} \right)^2} = 2/n^3 \end{aligned}$$

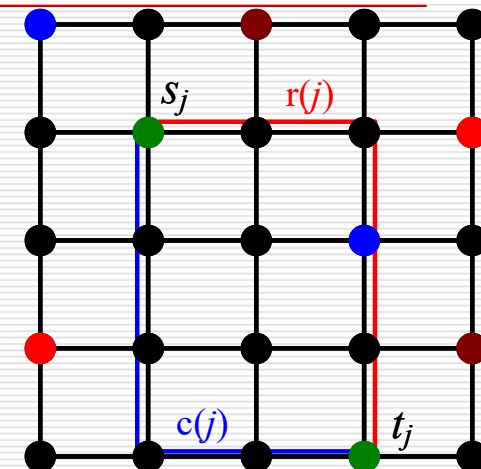
$$\underbrace{\frac{\mathbb{E}[X_j]}{n}}_{\frac{1}{4}} \cdot \frac{\overbrace{6\sqrt{\ln n}}^{\varepsilon}}{\sqrt{n}} = \frac{3}{2}\sqrt{n \ln n}$$

# 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  $A_1, \dots, A_k$ ,  $|A_j| = a_j n \geq \gamma n$ , 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  $A_j$ ,  $(1 - \varepsilon)a_j \leq \beta_j \leq (1 + \varepsilon)a_j$

# VLSI Routing

- Grid  $n \times n$  και  $k$  ζεύγη κορυφών  $(s_j, t_j)$  που πρέπει να **συνδέσουμε** με μονοπάτια.
  - **Δύο** μόνο **δυνατότητες** για κάθε ζεύγος  $j$ :  
 $r(j)$ : πρώτα ευθεία μετά κάθετα.  
 $c(j)$ : πρώτα κάθετα μετά ευθεία.
- Συνδέσεις που **ελαχιστοποιούν φορτίο** (#μονοπατιών) κάθε **ακμής**.
  - **NP-complete**. Εκφράζεται ως **Ακέραιο Γραμμικό Πρόγραμμα**:



$$\begin{aligned} & \min W \\ \text{s.t. } & \sum_{e \in r(j)} x_j + \sum_{e \in c(j)} (1 - x_j) \leq W \quad \forall e \in E \\ & x_j \in \{0, 1\} \quad \forall j \in [k] \end{aligned}$$

# VLSI Routing

---

- Λύνουμε σε **πολυωνυμικό χρόνο** το αντίστοιχο (μη Ακέραιο) **Γραμμικό Πρόγραμμα**:

- Βέλτιστη κλασματική λύση  $W^* \leq$  βέλτιστη ακέραια λύση.

$$\begin{aligned} & \min W \\ \text{s.t. } & \sum_{e \in r(j)} x_j + \sum_{e \in c(j)} (1 - x_j) \leq W \quad \forall e \in E \\ & x_j \geq 0 \quad \forall j \in [k] \end{aligned}$$

- Ντετερμινιστική στρογγυλοποίηση:
  - Για κάθε  $j$ , αν  $x_j^* \geq 1/2$  στην βέλτιστη ΓΠ-λύση,  $(s_j, t_j)$  συνδέεται με  $r(j)$ , διαφορετικά με  $c(j)$ .
  - Λόγος προσέγγισης 2, επειδή  $\max(x_j^*, 1 - x_j^*) \geq 1/2$ .



# VLSI Routing

---

## □ Randomized rounding:

- Για κάθε  $j$ ,  $(s_j, t_j)$  συνδέεται με  $r(j)$  με πιθανότητα  $x_j^*$ , διαφορετικά συνδέεται με  $c(j)$ .
- Τυχαία μετ/τη  $W$ : μέγιστο φορτίο ακμής στην (ακέραια) λύση  $(x_1, \dots, x_n)$  που προκύπτει.  $E[W_e] \leq W^*$ .
- Θέτουμε  $m = 2n(n-1)$  (#ακμών στο grid).
- Εφαρμόζοντας Chernoff bounds με  $\varepsilon = \sqrt{3 \ln(m/\delta) / W^*}$  έχουμε ότι αν  $W^* \geq 3 \ln(m/\delta)$ , τότε:

$$\Pr \left[ W \leq W^* + \sqrt{3W^* \ln(m/\delta)} \right] \geq 1 - \delta$$