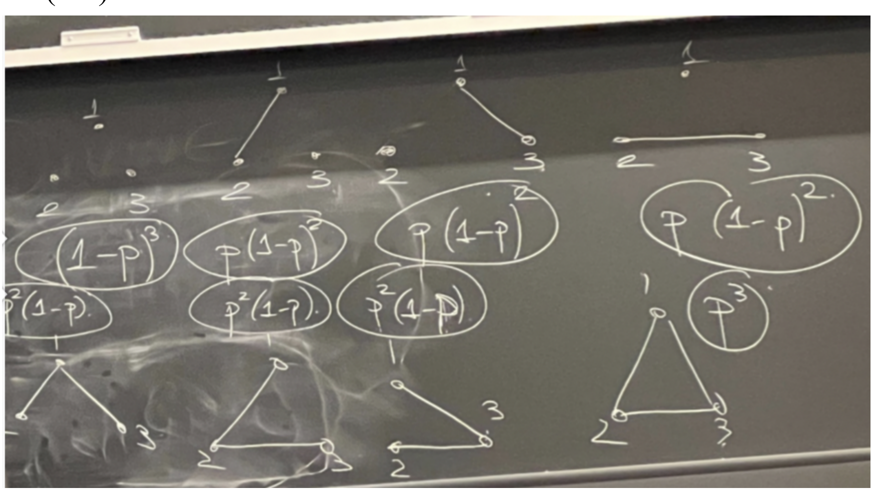
CAS CS 365
Lec 12

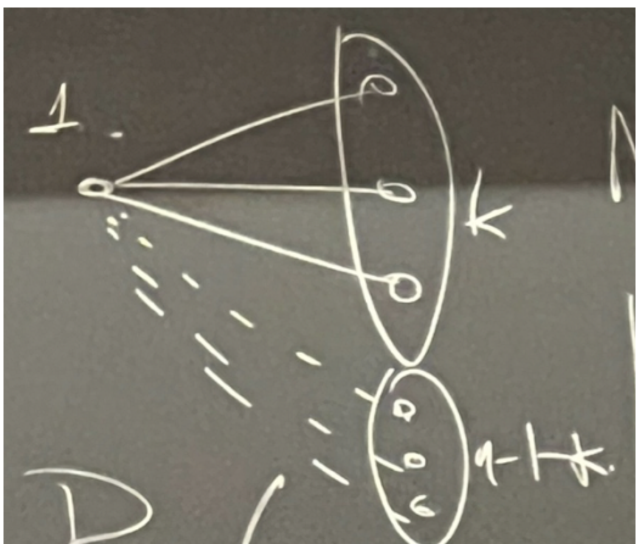
Graphs

1. Graph (or Networks)
    a. Model a wide variety of datasets
    b. Natural graph: social networks, collaboration networks
        i. Humans (every node is a human)
        ii. Relationship between humans (edge between them)
        iii. Lots of research on drug discovery graph neural networks → every node is a protein and if they interact, there is an edge
    c. Similarity graphs
        i. Two time series → edge if there is a similarity between two stocks
    d. Google
        i. Picture → nodes
        ii. Edge if there are similarity or relationship between two pictures
2. Erdos-Renyi Model (Random Binomial graphs)
    a. Graph requires vertex: V = {1,2,...,n} = [n]
    b. G(n, p)
        i. n → number of nodes
        ii. p → probability of connection (edge)
    c. For each pair of nodes, there are nC2 pairs and we toss a coin and with probability p, we add an edge
        i. For example, there are 3 nodes
            1. N nodes (1,...,n). How many possible graphs exist?
                a. 2 ^ (${}_nC_2$)

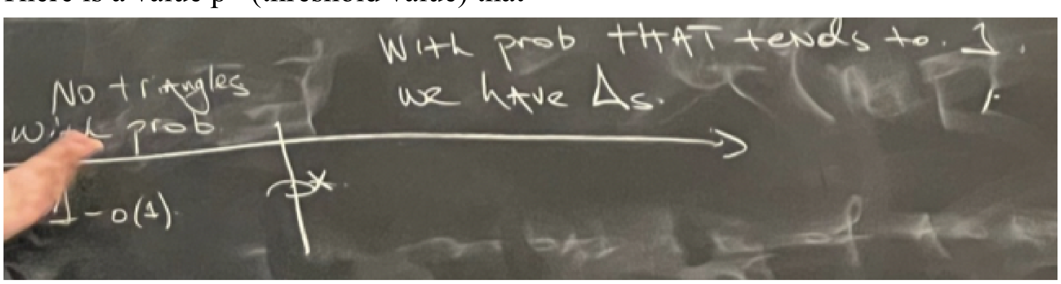


        ii. G(3,p)
            1. If all edges are disconnected, the probability is (1-p)^3
            2. If one edge is connected, the probability is p(1-p)^2

        3. If two edges are connected, the probability is p^2(1-p)
        4. If all edges are connected, the probability is p^3
    d. If there is a graph, there are n edges
        i. This means that there are ${}_nC_2$ - m non-edges
        ii. This means that the probability is p^m (1-p)^(${}_nC_2$ - m)
    e. What is the probability that a sample has exactly m edges
        i. P(a sample has exactly m edges) = (${}_nC_2$)C_m * p^m * (1-p)^(${}_nC_2$ -m)
    f. Number of edges in G(n,p) = M
    g. M = $X_{1,2}$ + $X_{1,3}$ + … + $X_{n-1,n}$
        i. Where Xa,b = 1 with probability p and 0 otherwise
    h. Pr(M=k) = (${}_nC_2$)C_k * p^k * (1-p)^(${}_nC_2$ -k)
3. Properties of graphs
    a. Neighbors of u = N(u) = {v: u and v are connected with an edge}
    b. |N(u)| = deg(u) = degree
        i. Degree of any node is Bin(n-1, p)
    c. Pr(deg(u) = k) = ${}_{n-1}C_k$ * p^k * (1-p)^(n-1-k)



4. Phase transition
    a. There is a value p* (threshold value) that



    b. T = number of triangles in G(n,p)
        i. Pr($\text{triangle}_{1,2,3}$)
        ii. E(T) = ${}_nC_3$ * p^3

5. Definition
    a. A node I is isolated if $deg(I) = 0$
    b. Let X be the random variable equal to the number of isolated nodes
        i. If there are 3 nodes, the expected number of isolated nodes is
            $3(1-p)^3 + p(1-p)^2 + p(1-p)^2 + p(1-p)^2$
            $= 3(1-p)^3+3p(1-p)^2 = 3(1-p)^2$
    c. Define $Y_i = 1$ if $deg(i) = 0$ and 0 otherwise
    d. Expected number of isolated nodes by linearity of expected value is
       $Pr(Y_i=1) = (1-p)^{(n-1)}$
       $E[x] = \sum$ i = 1 to n $E(Y_1) = n*(1-p)^{(n-1)}$
    e. Covariance of Yi, Yj
        i. $Cov(Y_i, Y_j) = E(Y_iY_j) - E(Y_i) E(Y_j)$
           $= E(Y_iY_j) - (1-p)^{2(n-1)}$
           $= (1-p)^{(2(n-2)+1)} - (1-p)^{2(n-1)}$
6. Definition
    a. $P = c\ln(n)/n \rightarrow$ if $n \rightarrow$ infinity, p goes to 0
    b. $E(x) = n(1-p)^{(n-1)}$
    c. $\lim(1+c/n)^n$ as $n \rightarrow$ infinity $= e^c$
    d. $\lim E[x]$ as $n \rightarrow$ infinity
       $= \lim n*(1-c\ln(n)/n)^{(n-1)}$ as $n \rightarrow$ infinity
       $= n * e^{(\ln(n-1)/n)}$ as $n \rightarrow$ infinity
       $= n^{(1-c)}$
       $\lim (1-c*\ln(n)/n)^{(n-1)}$ as $n \rightarrow$ infinity
       $=e^{(-c*\ln(n)/n*(n-1)}$ as $n \rightarrow$ infinity (as $(n-1)/n$ goes to infinity, it goes to 1)
       $=e^{(-c*\ln(n)/n)}$
       $=n^{-c}$
       Therefore, $n*n^{-c} = n^{(1-c)}$

        i. If $c > 1$, then $E[x] = n^{(1-c)} = n^{(-delta)} = o(1) \rightarrow$ goes to 0
            1. If c = 1 + delta (where delta is positive)
                a. $Pr(X >= 1) <= E[x] \rightarrow 0$
                b. $Pr(X = 0) \rightarrow 1 - o(1) = 1$ when $p > \ln(n)/n$