

Graph Mining Class Notes for Feb. 12, 2009

Distribution of Cliques in a Random Graph

Presented by Ruoming Jin. Transcribed by Victor Lee

Department of Computer Science, Kent State University, Kent, Ohio, U.S.A.,

1 Nonoverlapping Cliques

What is the probability that there are t 4-cliques (K_4) in a random graph $G_{n,p}$?

$$Pr(X = t) = Pr(X_1 + X_2 + \dots + X_c = t), c = \binom{n}{4}$$

and each X_i is a unique combination of 4 vertices in the graph. Then

$$X_i = \begin{cases} 0 & : 1 - p^6 \\ 1 & : p^6 \end{cases}$$

If cliques were independent (didn't overlap), we could use a simplified model where we simply consider the probability of t vertices being "cliques." For this, we can write a binomial distribution:

$$Pr(X = t) = \binom{n}{t} p^t (1 - p)^{n-t}$$

However, cliques can overlap, sharing vertices and edges, so the formulation is not that simple. We will show, however, that if n is large, we can describe the distribution as a Poisson distribution.

1.1 Poisson Distribution

The **Poisson distribution** describes the discrete probability that t independent events will occur in one unit of time, given an average event rate λ . It is used in queuing theory, to answer questions such as, How many customers might come to the check-out line at the store in one hour?

$$Pr_{poisson}(t, \lambda) = \frac{e^{-\lambda} \lambda^t}{t!}, \lambda > 0, t \geq 0 \quad (1)$$

Lemma 1. *The mean and variance for the Poisson distribution are both λ .*

Lemma 2. *The binomial distribution converges to the Poisson distribution when $n \rightarrow \infty$.*

Proof. Let $\lambda = np$. Note that $np = E(m)$, the expected number of edges.

$$\begin{aligned}
\lim_{n \rightarrow \infty} \Pr(X = t) &= \lim_{n \rightarrow \infty} \binom{n}{t} p^t (1-p)^{n-t} \\
&= \lim_{n \rightarrow \infty} \left(\frac{n!}{(n-t)!t!} \right) p^t (1-p)^{n-t} \\
&\approx \lim_{n \rightarrow \infty} \left(\frac{n^t}{t!} \right) \left(\frac{\lambda}{n} \right)^t \left(1 - \frac{\lambda}{n} \right)^n \left(1 - \frac{\lambda}{n} \right)^{-t} \\
&\approx \frac{\lambda^t}{t!} (e^{-\lambda}) (1)
\end{aligned}$$

□

Corollary 1. *If cliques are independent and n is large, the the Poisson distribution describes the clique distribution where $\lambda = \binom{n}{4} p^6$.*

Proof. In the previous lecture, we demonstrated that for 4-cliques, $EX = \binom{n}{4} p^6$. Combine Lemmas 1 and 2. □

2 Overlapping Cliques

$X = X_1 \wedge X_2 \wedge \dots \wedge X_m$, where X_i are random variables.

Viewed differently, $B = B_1 + B_2 + \dots + B_m$, where B_i are Bernoulli trials.

Consider: What is the probability that there are *no* cliques? According to the inclusion-exclusion principle,

$$\begin{aligned}
\Pr(X = 0) &= \Pr(X_1 = 0 \wedge X_2 = 0 \wedge \dots \wedge X_m = 0) \\
&= 1 - \sum_{i=1}^m \Pr(B_i) \\
&\quad + \sum_{i,j:1 \leq i < j \leq m} \Pr(B_i \wedge B_j) \\
&\quad - \sum_{i,j,k:1 \leq i < j < k \leq m} \Pr(B_i \wedge B_j \wedge B_k) \\
&\quad \dots + (-1)^r \sum \Pr(B_i \wedge B_j \wedge \dots \wedge B_{i+r})
\end{aligned}$$

Let $S^{(r)} = \sum_i \Pr(B_{i1} \wedge B_{i2} \wedge \dots \wedge B_{ir})$. Then we can rewrite

$$\Pr(X = 0) = 1 - S^{(1)} + S^{(2)} \dots + (-1)^r S^{(r)}$$

Lemma 3. *If S is defined as above and λ is a constant, then*

(1) $S^{(1)} = EX = \sum Pr(B_i) = \lambda$

(2) *For every fixed r ,*

$$S^{(r)} = \sum_i E(X_{i1}X_{i2} \cdots X_{ir}) \rightarrow \frac{\lambda^r}{r!}$$

We also claim that $P(X = 0) = e^{-\lambda}$ (proven in the next section) and $P(X = t)_{t>0} = \frac{e^{-\lambda}\lambda^t}{t!}$.

Proof sketch. Show that $S^{(2)} = \binom{n}{4} p^6$.

$$\begin{aligned} S^{(2)} &= \sum_{i=1}^{nC_4-1} \sum_{j=i+1}^{nC_4} Pr(X_i = 1 \wedge X_j = 1) \\ &= 1/2 \sum_{i=1}^{nC_4} \sum_{j=1, j \neq i}^{nC_4} Pr(X_i = 1 \wedge X_j = 1) \\ &= 1/2 \sum_i \sum_j Pr(X_j = 1 | X_i = 1) Pr(X_i = 1) \\ &= 1/2 \sum_i Pr(X_i = 1) \sum_j Pr(X_j = 1 | X_i = 1) \\ &= 1/2 \binom{n}{4} p^6 \sum_j Pr(X_j = 1 | X_i = 1) \end{aligned}$$

Break down $P(X_j = 1 | X_i = 1)$ into the four cases: (1) no vertices are shared; (2) one vertex is shared; (3) two vertices are shared; and (4) three vertices are shared.

$$S^{(2)} = 1/2 \binom{n}{4} p^6 \left[\binom{n-4}{4} p^6 + \binom{4}{1} \binom{n-4}{3} p^6 + \binom{4}{2} \binom{n-4}{2} p^5 + \binom{4}{3} \binom{n-4}{1} p^3 \right]$$

When n is large, we expect the first term to dominate. We should take a little care with p , however. $0 < p < 1$, so higher powers of p are smaller. If we assume that p is at or above the threshold $n^{-2/3}$, then we confirm that the first term dominates. So,

$$\begin{aligned} \lim_{n \rightarrow \infty} S^{(2)} &= 1/2 \binom{n}{4} p^6 \left[\binom{n}{4} p^6 \right] \\ &= \frac{\lambda^2}{2}, \text{ where } \lambda = \binom{n}{4} p^6 \end{aligned}$$

□

3 Proof for $Pr(X = 0) \rightarrow e^{-\lambda}$

Theorem 1. *For $t = 0$: $\lim_{r \rightarrow \infty} Pr(X = 0) = e^{-\lambda}$.*

Proof.

$$Pr(X = 0) = 1 - S^{(1)} + S^{(2)} \dots + (-1)^r S^{(r)}$$

We will find lower and upper bounds for $Pr(X = 0)$ and then show that as $n \rightarrow \infty$, the bounds become $e^{-\lambda} \pm \epsilon$, for an arbitrarily small ϵ .

$$\begin{aligned} Pr\left(\bigcup_{i=1}^n E_i\right) &\leq \sum_{i=1}^n Pr(E_i) \text{ [Bonferroni's inequality]} \\ &\leq \sum_i Pr(E_i) - \sum_{i < j} Pr(E_i \cap E_j) \dots + (-1)^{2s} \sum_{i < j < \dots < 2s} Pr(E_i \cap E_j \cap \dots \cap E_{2s}) \end{aligned}$$

because the last term is positive. Similarly,

$$Pr\left(\bigcup_{i=1}^n E_i\right) \geq \sum_i Pr(E_i) - \sum_{i < j} Pr(E_i \cap E_j) \dots + (-1)^{2s+1} \sum_{i < j < \dots < 2s+1} Pr(E_i \cap E_j \cap \dots \cap E_{2s+1})$$

because the last term is negative. Using the earlier definition of S , we can rewrite the inequalities:

$$\begin{aligned} Pr\left(\bigcup_{i=1}^n E_i\right) &= 1 - S^{(1)} + S^{(2)} \dots + (-1)^n S^{(n)} \\ \sum_{r=1}^{2s+1} (-1)^r S^{(r)} &\leq Pr(X = 0) \leq \sum_{r=1}^{2s} (-1)^r S^{(r)} \end{aligned}$$

Choose n such that $0 \leq r \leq 2s$. Now, define two inequalities which will help us later:

(A) Since $\lim_{r \rightarrow \infty} S^{(r)} = \frac{\lambda^r}{r!}$,

$$\left| S^{(r)} - \frac{\lambda^r}{r!} \right| < \frac{\epsilon}{2(2s+1)}$$

(B) Since $\sum_{n=1}^{\infty} (-1)^n \frac{x^n}{n!} = e^{-x}$,

$$\left| \sum_{r=0}^{2s} (-1)^r \frac{\lambda^r}{r!} - e^{-\lambda} \right| < \epsilon/2$$

From (A), $S^{(r)} \leq \frac{\lambda^r}{r!} + \frac{\epsilon}{2(2s+1)}$

Multiply both sides by $(-1)^r$ and sum for $r = 0$ to $2s$. Note that for odd values of r , we are changing the sign. So, the inequality for the sum only holds if the even terms out-weigh the odd terms. We do not prove this, but we can observe that summing

from 0 to $2s$, there is one more even term than odd term.

$$\begin{aligned}
\sum_{r=0}^{2s} (-1)^r S^{(r)} &\leq \sum_{r=0}^{2s} (-1)^r \left(\frac{\lambda^r}{r!} + \frac{\epsilon}{2(2s+1)} \right) \\
&\leq \sum_{r=0}^{2s} (-1)^r \frac{\lambda^r}{r!} + \sum_{r=0}^{2s} \frac{\epsilon}{2(2s+1)} \\
&= \sum_{r=0}^{2s} (-1)^r \frac{\lambda^r}{r!} + \epsilon/2 \\
&\leq \left(e^{-\lambda} + \epsilon/2 \right) + \epsilon/2 \text{ [From inequality (B)]} \\
&= e^{-\lambda} + \epsilon
\end{aligned}$$

So, the upper bound for $\Pr(X \leq 0)$ is $e^{-\lambda} + \epsilon$, for arbitrarily small ϵ . Presumably, a similar proof can be used to show that the lower bound is $e^{-\lambda} - \epsilon$. \square