Q) How to determine how efficient Bloom filters are?

→ Say there are "n" sets who's membership is to be represented using a bloom filter of "m" bits, with a false +ve rate of ϵ

→ "p" of a single bit not set any time : $\left(1-\frac{1}{m}\right)$

> Assuming the hash fns are very random

→ "p" of none of the m bits are set : $\left(1-\frac{1}{m}\right) \times \left(1-\frac{1}{m}\right) \times \cdots$

$= \left(1-\frac{1}{m}\right)^n$

→ when there are k hash fns : $\left(1-\frac{1}{m}\right)^{nk}$

→ when we test for an element that doesn't exist k hashes "p" of all of those k bits to be set : "p" is false +ve

$$p = \left(1-\left(1-\frac{1}{m}\right)^{nk}\right)^k \quad \text{— (1)}$$

→ let's approximate m being very large to find true efficiency of bloom filters

$$\lim_{m\to\infty} \left(1-\frac{1}{m}\right)^k$$

$$= \lim_{m\to\infty} \left(1+\frac{(-1)}{m}\right)^k \quad \text{— (2)}$$

for real world, the approximation of (1) = (2) is close enough ∴

$$\Rightarrow \lim_{m\to\infty} \left(1+\frac{(-1)}{m}\right)^k = \underline{\underline{e^{-1}}}$$

Using :

$$\left[\lim_{n\to\infty} \left(1+\frac{\alpha}{n}\right)^n = e^{\alpha}\right]$$

$$\left[\lim_{m\to\infty} \left(1-\frac{1}{m}\right)^m = e^{-1}\right]$$

└ (1)

$$\left(1-\frac{1}{m}\right)^k = \left(\left(1-\frac{1}{m}\right)^m\right)^{k/m} \approx e^{-k/m} \approx \left(1-\frac{1}{m}\right)^k \quad \text{— (3)}$$

Sub (3) in (0)

$$P = \left(1-e^{-\frac{kn}{m}}\right)^k$$

$$P(k) = e^{\ln\left(\left(1-e^{-\frac{kn}{m}}\right)^k\right)} \Rightarrow P(k) = e^{k\ln\left(1-e^{-kn/m}\right)}$$

> lets differentiate the power of e. to make it easier to differentiate against k as the minima remains same. (they have the same x intercept anyways)

$$\frac{d\left[k\ln\left(1-e^{-kn/m}\right)\right]}{dk} = \ln\left(1-e^{-\frac{kn}{m}}\right) + k\,\frac{\frac{n}{m}\times e^{-\frac{kn}{m}}}{1-e^{-\frac{kn}{m}}} \quad \text{— (4)}$$

let $e^{-\frac{kn}{m}} = q_1 \Rightarrow \ln(q_1) = -\frac{kn}{m}$

(4) $\Rightarrow \ln(1-q_1) - \ln(q_1)\frac{q_1}{1-q_1} = 0$ (eq to 0 to find min)

$\Rightarrow q_1 = 1/2 \Rightarrow e^{-kn/m} = 1/2$

$$-\frac{kn}{m} = \ln(1/2)$$

$$\boxed{k = \ln(2)\frac{m}{n}} \quad \text{— (f1)}$$

false +ve rate $= \epsilon = \left(1-e^{-\frac{kn}{m}}\right)^k$ (from eq (0))

LHS

$$= \left(1-e^{-\ln(2)\frac{m}{n}\times\frac{n}{m}}\right)^{\ln(2)\frac{m}{n}}$$

$$= \left(1-e^{-\ln(2)}\right)^{\frac{\ln(2)m}{n}} = \left(1-e^{\ln(1/2)}\right)^{\frac{\ln(2)m}{n}}$$

$$= \left(1-1/2\right)^{\frac{\ln(2)m}{n}} = \left(1/2\right)^{\frac{\ln(2)m}{n}}$$

RHS

$$\varepsilon = \left(\frac{1}{2}\right)^{\log_{1/2}(\varepsilon)} \Rightarrow$$

(LHS) (RHS)

$$\left(\frac{1}{2}\right)^{\log_{1/2}(\varepsilon)} = \left(\frac{1}{2}\right)^{\frac{\ln(2)\,m}{n}}$$

$$\log_{1/2}(\varepsilon) = \frac{\ln(2)\,m}{n} \Rightarrow \frac{m}{n} = \frac{\log_{1/2}(\varepsilon)}{\ln(2)}$$

$$\log_{1/2}(\varepsilon) = \ln(2)\,\frac{m}{n}$$

$$\frac{\log_2(\varepsilon)}{\log_2(1/2)} = \ln(2)\,m/n$$

$$(-1)\quad \log_2(\varepsilon) = \ln(2)\,m/n$$

$$m/n = -1.44\log_2(\varepsilon) \quad \leftarrow \quad (f2)$$

So, to decipher what (f2) is saying, Say we want to rep n sets using m bits:

$$m = n \times 1.44 \times \log_2(\varepsilon)$$

You would need $1.44 \times \log_2(\varepsilon)$ # of bits for each member of set of n, where the ideal minima would be $1 \times \log_2(\varepsilon)$ for each member of n, where ε is the false +ve rate.

IDEALLY: $\log_2(\text{rate})$ would be the # of required bits.