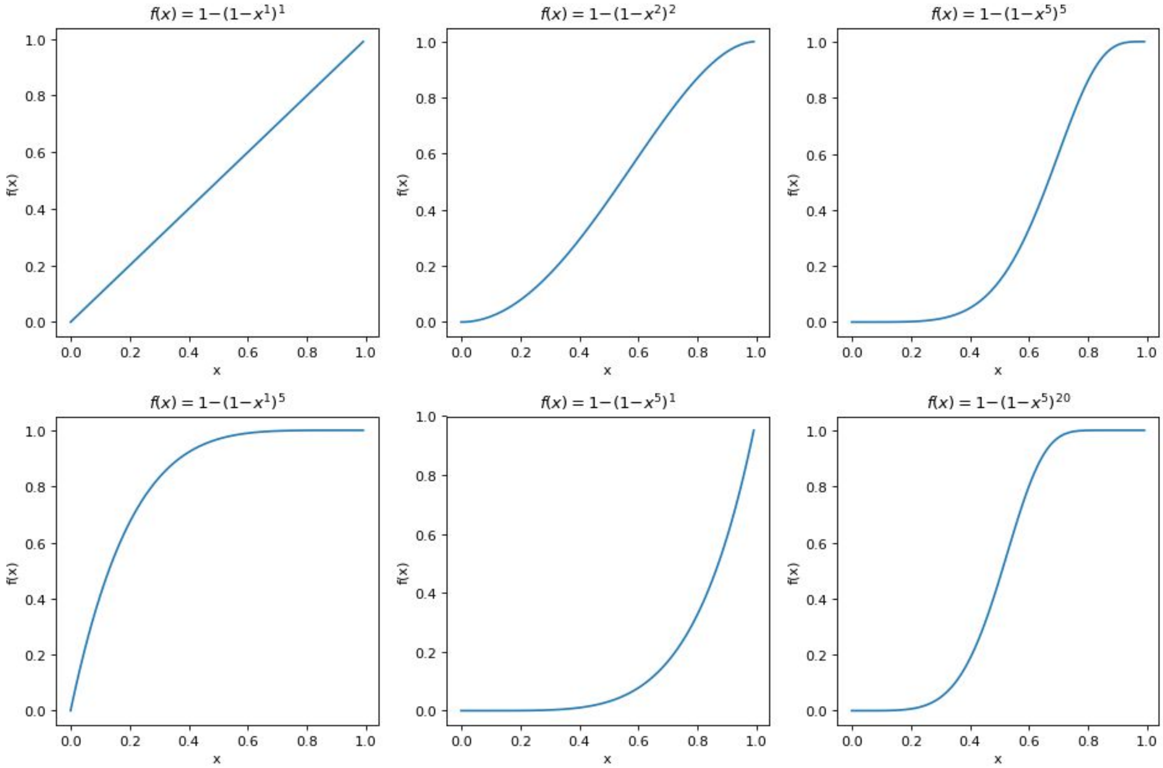
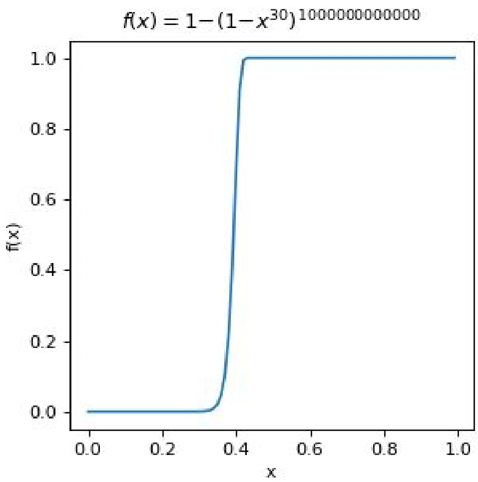
Assignment 1
CS-UH-2218: Algorithmic Foundations of Data Science
Simon Seo (ms9144), Khaled Al Hosani (kah579)

Problem 1

$$f(x) = 1 - (1 - x^r)^b$$



$$f(x) = 1 - (1 - x^{30})^{10^{12}}$$
$$f'(0.4) = 27.30$$



Problem 2 (Exercise 1.2.1)

Default:
Population ***p*** = 10^9 people
Frequency ***f*** = 1 visit / (100 days * person)
Hotel size ***s*** = 100 people / hotel
hotels ***h*** = 10^5 hotels
Observation period ***o*** = 10^3 days
Report threshold ***t*** = 2 days (in which 2 people were in same hotel)

# of suspected pairs = (chance of two people visiting same hotels on two/three days) * (total # of possible pairs/triplets of days) * (total # of possible pairs of people)

$$\#\text{suspected pairs} = \left(\frac{f^2}{h}\right)^t \times \binom{o}{t} \times \binom{p}{2}$$

a) Observation period ***o*** = 2000 days

$$\#\text{suspected pairs} = \left(\frac{(10^{-2})^2}{10^5}\right)^2 \times \binom{2\times10^3}{2} \times \binom{10^9}{2}$$
$$\approx 10^{-18} \times 2\times10^6 \times 5\times10^{17}$$
$$= 100{,}000 \text{ pairs}$$

b) # hotels ***h*** = 2*10^5 hotels, population ***p*** = 2*10^9 people

$$\#\text{suspected pairs} = \left(\frac{(10^{-2})^2}{2\times10^5}\right)^2 \times \binom{10^3}{2} \times \binom{2\times10^9}{2}$$
$$\approx \frac{1}{4}\times10^{-18} \times 5\times10^5 \times \cancel{5\times10^{17}}\ 2\times10^{18}$$
$$\approx 250{,}000 \text{ pairs.}$$

c) Report threshold ***t*** = 3 days

$$\#\text{suspected pairs} = \left(\frac{(10^{-2})^2}{10^5}\right)^3 \times \binom{10^3}{3} \times \binom{10^9}{2}$$
$$\approx 10^{-27} \times \frac{5}{3}\times10^8 \times 5\times10^{17}$$
$$= 1/12 \text{ pairs}$$

Since the expected number of suspected pairs is 1/12, there would be approximately 1 pair of suspicious people every 12 observation cycles.

Problem 3 (Exercise 2.3.1)

(a) The largest integer.

map: for each integer x_i produce key-value pair $(1, x_i)$
combiner: take list of KV pairs $[(1,x_1),(1,x_2),\cdots]$
and produce KV pair $(1, \max_{i\le n}(x_i))$
reduce: with list $(1, [x_{max1}, x_{max2}, x_{max3}\cdots])$
produce $(1, \max_{i\le m}(x_{max_i}))$
where n is number of integers in mapper task,
m is number of mappers
output: $\max_{i\le m}(x_{max_i})$ (max of individual maxima)

(b) The average of all the integers.

map: $x_i \mapsto (1, x_i)$
combiner: $[(1,x_1),(1,x_2),\cdots] \mapsto (1,(\sum^n 1, \sum^n x_i))$
reducer: $(1, [(n_1,S_1),(n_2,S_2),\cdots]) \mapsto (1, (\sum^m n_i, \sum^m S_i))$
where n_i = # of integers in mapper i
S_i = sum of integers from mapper i
m = # of mappers.
output: $\sum^m S_i / \sum^m n_i$ (total sum / total count)

(c) The same set of integers, but with each integer appearing only once.

map: $x_i \mapsto (x_i, 1)$
reduce: $(x_i, [1,1,1,\cdots]) \mapsto (x_i, 1)$
output: x_i (list of keys from reducer)

(d) The count of the number of distinct integers in the input.

From outputs of (c), take list of unique integers as input.
map2: int[] x $\mapsto$ (1, count of x_i's)
reduce2: $(1, [c_1, c_2, c_3\cdots]) \mapsto (1, \sum c_i)$
where c_i is count of unique integers from each mapper
Output: $\sum c_i$