EE412 Foundation of Big Data Analytics, Fall 2021

HW1

Name: 권혁태

Student ID: 20180036

Discussion Group (People with whom you discussed ideas used in your answers):

김기영

On-line or hardcopy documents used as part of your answers:

## Answer to Problem 1

Explanation about my algorithm

주어진 데이터를 보게 되면, 모든 사람들에 대해서 친구 리스트를 주고 있습니다.

0번 사람에 대한 친구가 1이 존재한다면, 1번 사람에 대해 친구로 0번이 들어가 있습니다.

제가 접근한 방법은 potential friend일 수 있는 candidate set을 만들고, 기존 데이터에서 만들 수 있는 friend set 전부를 제거하는 방식입니다.

Potential friend일 수 있는 candidate set을 만드는 방법은 다음과 같습니다.

0번 사람에 대해 친구 1,2,3이 있다고 생각할 때, 우리는 1,2,3에 대해서 (1,2), (2,3), (1,3)이 서로 친구일 가능성이 있습니다. 이들은 모두 0번 사람을 알고 있기 때문입니다. 따라서 spark에서는 line을 읽으면서 각 line의 friend list를 정렬한 다음, 위와 같은 pair를 만들어 발행하는 식으로 candidate를 생성합니다.

하지만 이 때, 1번 사람에 대한 줄에서 2번 사람이 친구로 나오는 것과 같은 경우가 존재합니다. 즉, 원래 친구인 경우를 제거해줘야합니다. 원래 친구인 집합을 originals라 하겠습니다. Originals를 만들기위해 각 라인을 읽으면서 key값과 friend list에서 element로 이루어진 pair를 만들어 발행합니다.

이후 전체 candidate rdd에서 originals를 subtract 해준 후, reduce를 써서 같은 rdd의 개수를 합치면 됩니다. Reduce해서 합쳤을 때, potential friends의 공통 지인이 나오는 이유는, candidate을 만들 때 예를 들어 0번 사람에 대해 1,2가 친구 리스트로 있지만, 나중에 10번 사람에 대해서도 1,2가 친구 리스트로 있다면 중복되서 rdd가 발행될 것이고 결국 같은 키를 가지고 있는 rdd의 개수를 새면 potential friends의 공통 지인 숫자가 나오게 됩니다.

Program elapsed time

3min 30s

* start: 23/10/05 16:50:03
* end: 23/10/05 16:53:37

## Answer to Problem 2

a)

For triangular matrix:

* N \* (N-1) / 2 \* 4

Frequent item이 N개 이므로, triangular matrix에서는 frequent item에 대해서 pair를 인덱스로 표시해서 값을 세기 때문에 위와 같이 나오게 됩니다.

For triples method

* 12 \* M

2M pairs가 하나 이상 존재하고, 이 중 M개 만이 frequent item을 포함하기 때문에 이들이triple method를 위한 candidate이 됩니다. 이 candidate에 대해서 hash table을 만들어야하기 때문에 12 \* M입니다.

b)

elapsed Time: 1090s

c)

elapsed Time: 1200s

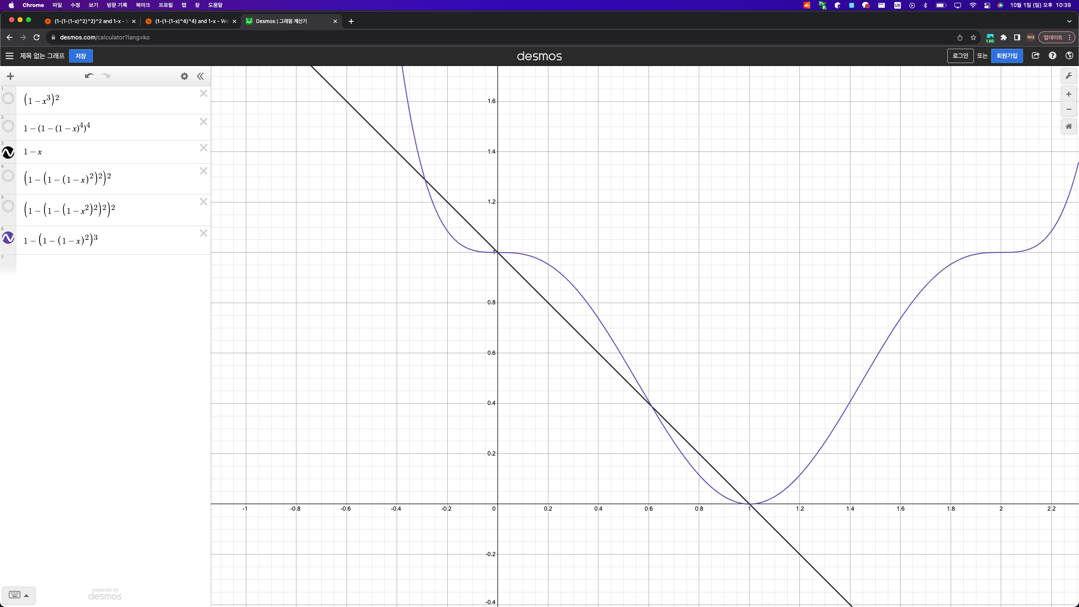
## Answer to Problem 3

a) all graph’s x axis is distance

- A 2-way AND construction followed by a 3-way OR construction.

(d1 , d2 , 1 – (1-(1- d1 )2 )3 , 1 – (1-(1- d2 )2 )3

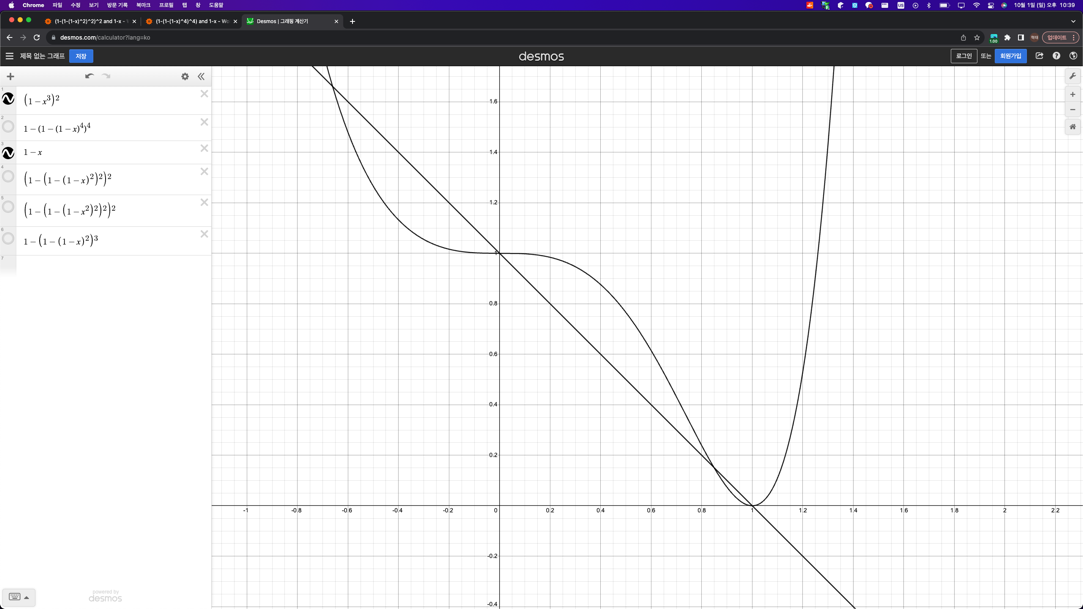
Increase high probability, decrease lower probability



- A 3-way OR construction followed by a 2-way AND construction.

( d1 , d2 , (1-(d1)3 )2 , (1- (d2)3 )2 )

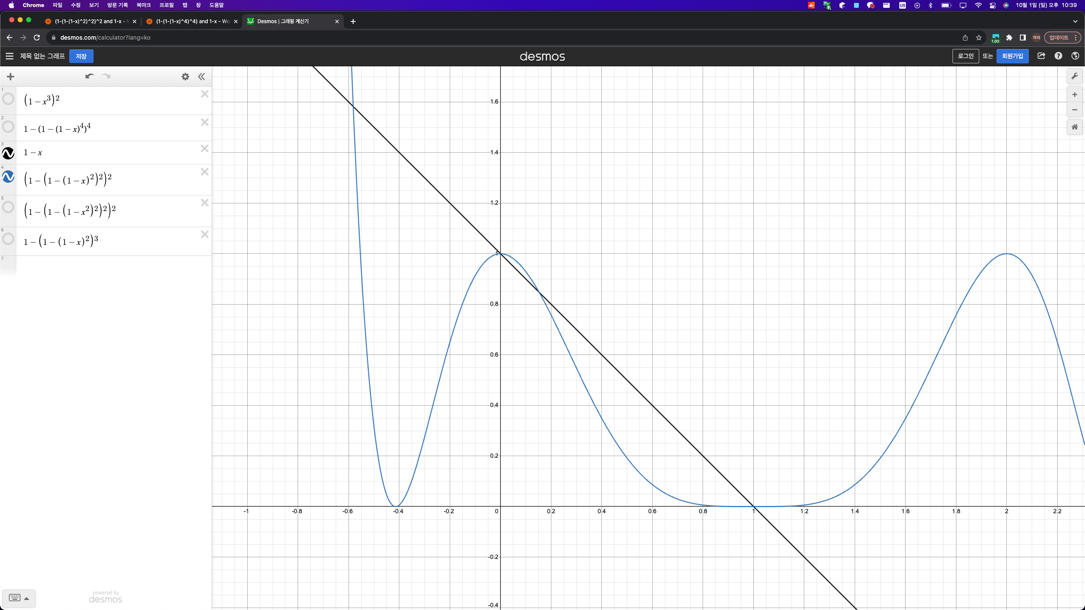
Increase high probability, increase low probability (when d2 is lower than 0.84)



- A 2-way AND construction followed by a 2-way OR construction, followed by a 2-way AND construction.

( d1 , d2 , (1 – (1 – (1-d1)2 )2 )2 , (1 – (1 – (1-d2)2)2 )2 )

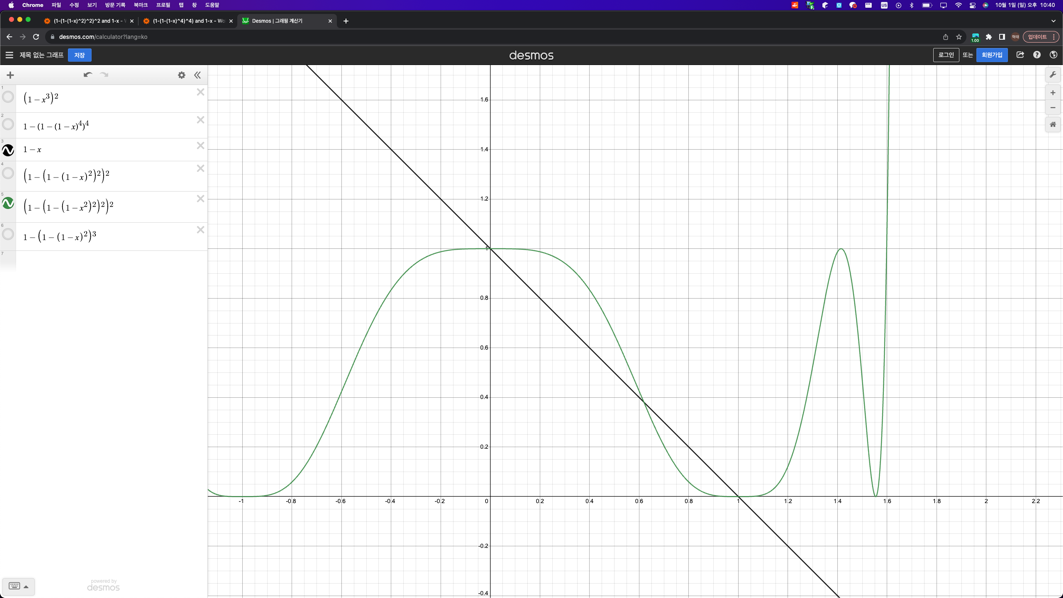
Decrease high probability, decrease low probability ( d1 is higher than 0.15 )



- A 2-way OR construction followed by a 2-way AND construction, followed by a 2-way OR construction followed by a 2-way AND construction.

(d1, d2 , (1 – ( 1 – ( 1- (d1)2 )2 )2 )2 , (1 – ( 1 – ( 1- (d2)2 )2 )2 )2 )

Increase high probability, decrease low probability



b)

elapsed time: 750s