How much is a beer ?

Richard GUILLEMOT

Thursday 10th October 2013

Algorithms & Data Challenge Berlin

Agenda

1. Business Model
2. Modeling and Simulation
3. Question to A&DC Berlin ?

1 – Business Model

- A mobile application to inform users of the price of the beer around.
- Difficulties : collect all the prices (costly and time-consuming)
- Solution : user collect the price and get a beer for free.
  http://coolmaterial.com/roundup/brilliant-beer-apps/

1 – Business Model

- Naïve assumption : all the users are reliable and honest.

$$\text{Total Cost} = \sum_{i=1}^{N} P_i$$

$\text{N} : \text{the number of Bars}$

$\text{P}_\text{i} : \text{the price of a beer ine the bar n° i}$

- Unfortunately all the users are not reliable and honest.

2 - Modeling and Simulation

- Our objective : Estimate more realistically the cost of the collection.

- Our proposal : Simulate the behavior of our users to estimate more realistically the cost of the collection.

2 - Modeling and Simulation

- Bars in the city: a matrix of

  N_B=70*70 = 4900 bars
- The price of a beer is taking 11 possible values:

  2.50 , 2.60… 3.00 ,… ,3.40,3.50

2 - Modeling & Simulation

We are assuming we have 4 type of users:

1. The “reliable normal” users.
2. The “not so reliable normal” users.
3. The “cheaters”.
4. The “good guys”.

2 - Modeling & Simulation

Organization of the collection:

- The process is dividing in round.
- For each round all our users will send us price of a beer in a bar.
- After each step, we and our users can access the result of the collection during the previous round.

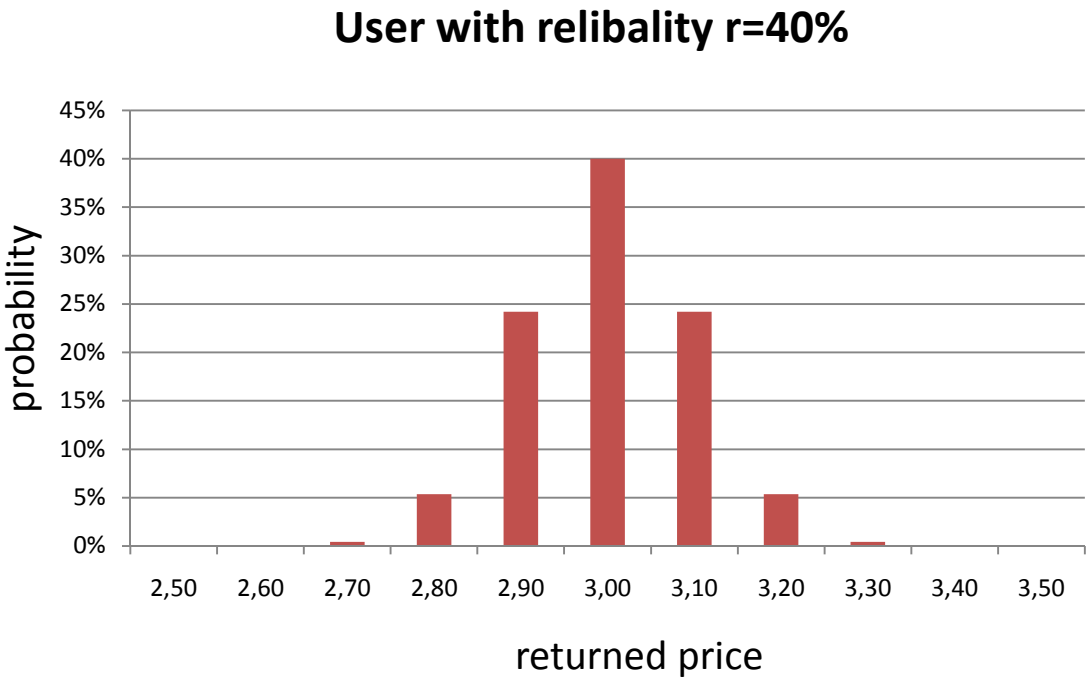
2 - Modeling & Simulation

Assumptions:

- There are less users as bars. $N_U < N_B$.
- Each user send us a price of single bar during a round (There is no collision between user).
- For each round all the users send us information.

2 - Modeling & Simulation

Response from a user and its reliability



2 - Modeling & Simulation

The reliability is the probability to return the good price with 5 cents tolerance.

$$G \sim N(p, \sigma)$$

$$P[p - 0.05 \leq G \leq p + 0.05] = r$$

$$\sigma\,?$$

2 - Modeling & Simulation

To improve the reliability of the global system we average the prices returned by different users.

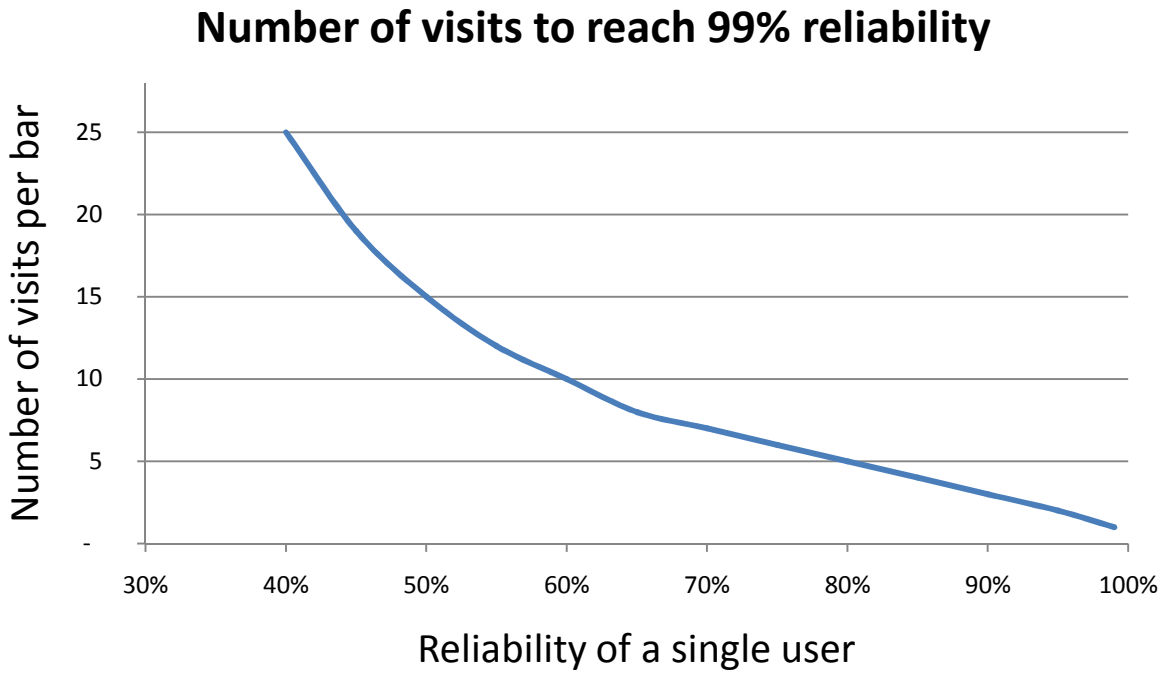
$$G_i \sim N(p, \sigma), 1 \leq i \leq N$$

$$\frac{1}{N}\sum_{i=1}^{N} G_i \sim N(p, \frac{\sigma}{\sqrt{N}})$$

$$r_{eq} \, ?$$

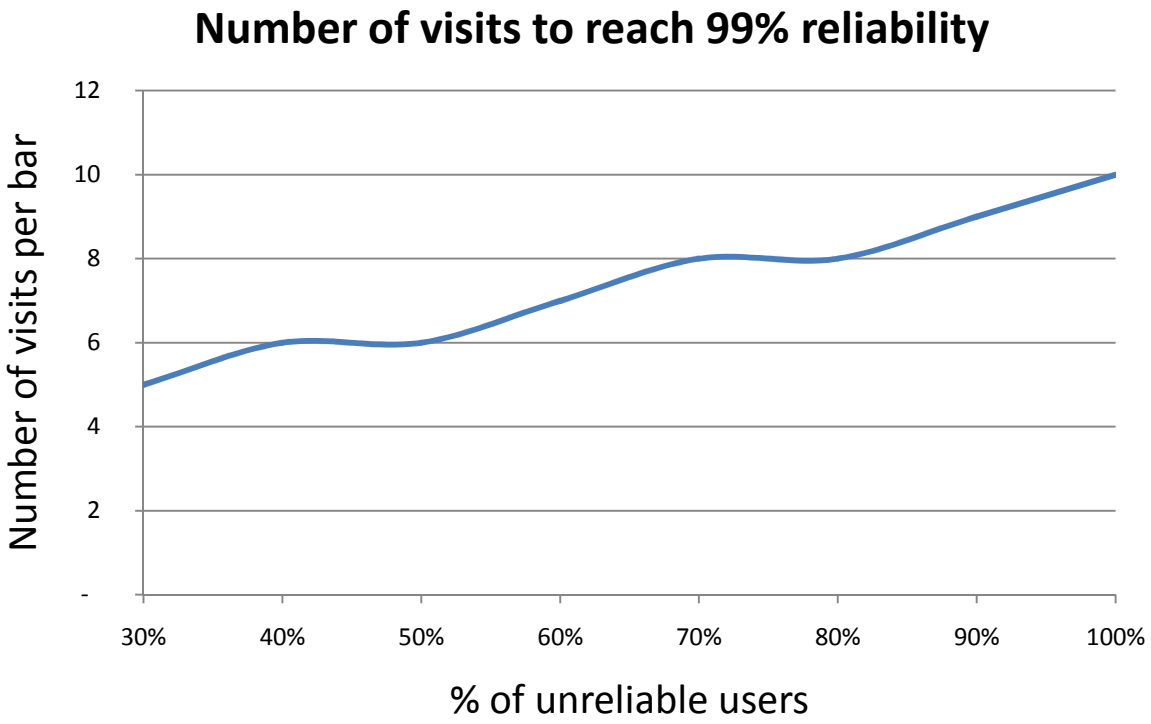
2 - Modeling & Simulation

Only reliable users (r_1=90%)



2 - Modeling & Simulation

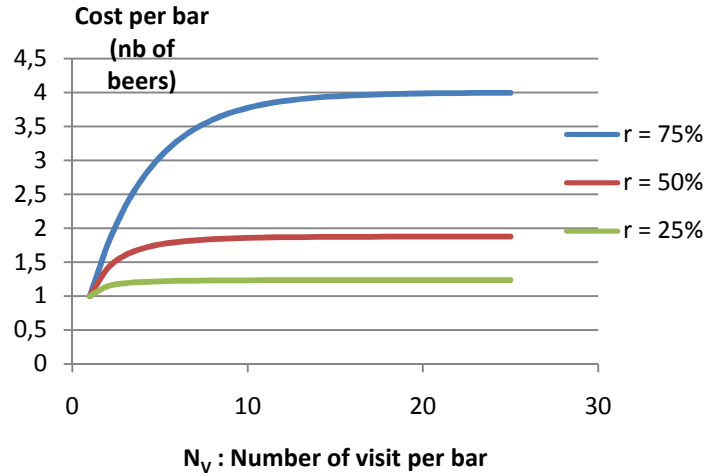
Reliable user (r_1=90%) Unreliable user (r_2=60%)



2 - Modeling and simulation

We repay the price of the beer to the first visitor and then a fraction (f) for the second, and a fraction of the fraction for the third one …

$$\text{Cost per bar} = P\frac{f^{N_V+1}-1}{f-1}$$



N_V : the number of visit per bar

f : the fraction of the price paid after the first visit

2 - Modeling & Simulation

The “cheaters”:

- They send the information available (already collected by the previous users).
- They are looking for the bars, where they get the most money (most expensive beer and the less visited).

2 – Modeling & Simulation

The “good guys”:

- They are 100% (r=100%) reliable. They return exactly the price of a beer.

- They are looking for the bars, where they get the most money (most expensive beer and the less visited).

2 - Modeling & Simulation

Consequences: each bar will be randomly visited by the normal users and not totally randomly by the cheaters and the good guys.

We assume the choice of the “normal” users is totally random (no localization, no preference, no earning expectation)

All the user behave independently.

2 - Modeling & Simulation

The cost of the collect will be “mostly” a function of those parameters:

Parameters		Parameters	
N_U	Number of users	P_{Good}	% of good guys
N_B	Number of bars	f	Fraction repaid after the first visit
N_R	Number of round	p_1	Reliability of the “reliable”
P_{Unrel}	% of unreliable users	P_2	Reliability of the “unreliable”
$P_{Cheaters}$	% of cheaters		

2 - Modeling & Simulation

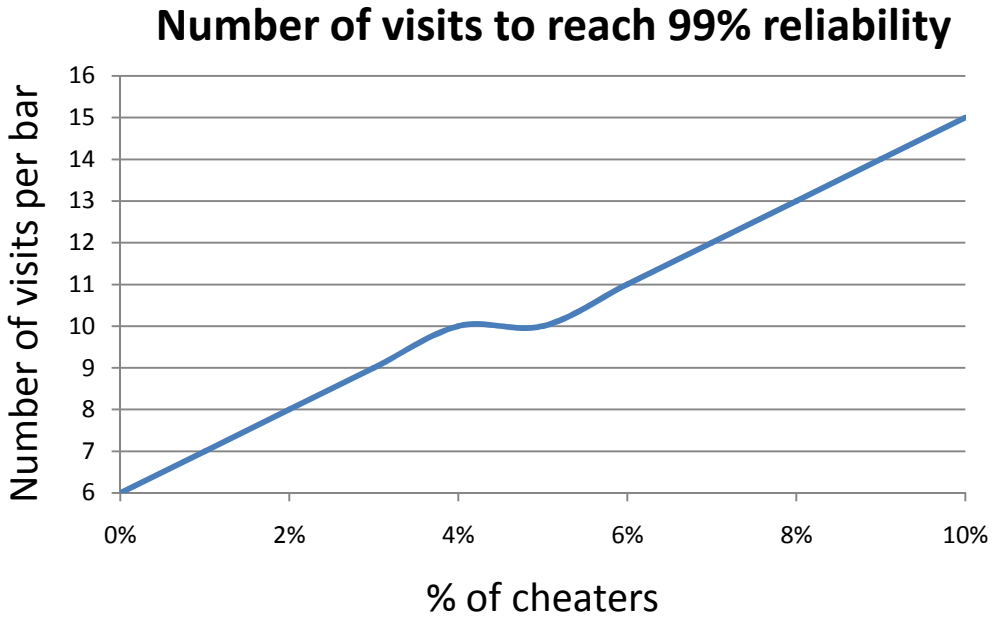
NBX	70
NBY	70
NU	4900
NR	7
pnorm	59%
punrel	40%
pcheat	1%
pgood	0%
f	50%
rreal	90%
runreal	60%
nosimul	1,00
datatype	Prices

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	3,3	2,6	3,4	3,3	2,6	3,5	3,4	2,7	3,1	2,8	2,6	3,0	2,8	2,7	3,0	3,5	3,5	3,5	3,5	3,5
2	3,3	3,3	3,2	2,6	2,8	3,3	3,5	3,0	2,5	3,2	2,9	2,6	2,9	2,7	3,3	2,6	3,3	2,5	2,7	2,9
3	3,3	3,5	2,8	2,8	3,3	2,8	2,7	2,5	3,4	2,7	2,9	3,4	2,7	3,3	2,8	3,5	3,1	3,3	3,0	3,5
4	3,1	3,0	2,8	2,8	3,2	3,0	3,2	3,2	3,2	3,5	3,0	3,2	2,6	3,4	2,7	3,0	3,4	3,1	2,7	3,0
5	3,4	3,0	3,1	2,9	2,9	2,7	3,0	2,6	2,9	2,9	2,6	3,3	2,7	2,5	2,6	2,8	2,7	3,0	2,7	3,3
6	2,7	3,3	3,1	2,8	3,2	2,6	3,1	3,0	3,0	2,9	3,0	3,2	2,8	2,8	3,2	3,1	2,7	3,3	3,2	2,8
7	2,7	3,0	2,8	2,5	3,0	3,4	2,7	2,7	3,3	3,4	2,7	2,9	2,7	2,6	2,7	2,7	2,7	3,3	2,9	3,0
8	3,2	2,6	3,0	2,8	3,1	2,8	2,7	3,5	3,0	2,7	3,5	3,1	3,0	2,7	3,0	2,7	2,7	2,9	3,0	2,9
9	2,5	2,9	3,2	3,2	3,0	2,8	3,0	2,9	3,4	2,8	3,1	3,4	3,1	3,0	3,4	3,1	3,3	3,0	3,1	3,0
10	3,0	2,9	2,9	2,7	3,3	3,0	2,6	3,4	2,6	2,6	2,7	3,2	2,9	3,1	3,3	2,7	3,3	2,9	2,6	2,8
11	2,8	2,8	3,2	2,6	3,2	3,2	3,0	2,7	3,2	3,3	3,2	2,7	2,7	2,8	2,6	3,4	3,5	3,2	2,7	3,2
12	3,5	3,4	3,2	3,3	2,8	2,7	3,1	3,4	2,6	2,9	3,4	3,3	3,4	2,8	3,3	2,8	2,8	3,2	3,1	3,1
13	3,3	3,0	2,9	3,2	2,6	2,6	2,8	2,6	2,7	3,0	2,8	2,8	2,9	3,0	3,0	2,6	3,0	2,6	3,4	2,6
14	3,3	3,5	2,8	2,8	3,3	2,7	3,0	2,7	2,5	3,3	2,7	3,0	3,2	2,9	3,0	2,7	2,7	3,0	2,8	3,2
15	3,0	3,0	3,2	3,2	3,3	3,2	2,9	2,6	3,2	2,7	2,9	2,9	3,3	2,5	3,3	2,6	2,8	2,8	3,1	2,8
16	3,3	2,9	3,4	3,3	3,5	2,9	2,7	3,3	2,6	3,5	3,2	3,3	2,6	3,3	3,0	2,9	3,0	2,6	3,4	3,5
17	3,3	3,0	2,9	3,3	3,4	3,0	2,9	3,4	3,3	3,1	2,9	3,4	3,3	2,5	3,3	3,2	2,9	3,1	2,7	2,6
18	3,0	3,0	2,7	3,2	3,0	2,7	2,6	2,7	2,6	2,6	3,4	2,8	3,1	2,7	3,4	2,5	3,2	3,3	3,1	2,7
19	3,3	3,1	3,1	3,1	3,4	2,6	3,1	3,4	2,5	2,9	2,6	2,5	3,4	3,1	3,0	2,9	3,3	2,6	2,7	3,1
20	2,7	3,4	3,1	2,9	3,1	2,9	2,9	2,9	2,7	3,2	3,4	3,3	2,6	3,5	2,6	2,6	2,6	3,2	2,7	2,7

2 - Modeling & Simulation

40% of unreliable users (r_1=90%)

The rest with normal users (r_2=60%)and cheaters



2 – Modeling & Simulation

All the information of the collection can be summarize through the following information:

$\mathrm{B}_{\mathrm{i,j}} = k$, the user i on the round j visit the bar k

$\mathrm{P}_{\mathrm{i,j}} = p$, he gives the price p

$1 \leq i \leq N_U, 1 \leq j \leq N_R, 1 \leq k \leq N_B$

3 – A&DC Question ?

Our approach is quite passive: based on a model we are expecting a cost.

Could we use a data analyze algorithm, which will recognize each category of users and help us to reduce the cost of collection by ignoring information send by cheaters and unreliable?

We could use the previous model to compare the cost reduction with the proposed algorithms.

3 – A&DC Question ?

Issues with the algorithm:

- How long will it take to learn ?
- Can we estimate its error rate ?
- Is it robust (to the behavior of the cheaters) ?