How Differential Privacy should Inform Randomized Response (and vice-versa)

By: Joseph “Joey” Knightbrook (knightbrook@google.com)
15 April, 2022

TL;DR

The field of differential privacy should learn from the social science technique of randomized response. Randomized Response is a technique in surveys to inject noise to ensure data privacy and get more honest answers to sensitive questions. I will show evidence that the strongest randomized response methods are the ones with a stronger differential privacy guarantee. And further I show evidence from a randomized response study where users were roughly as willing to answer a sensitive question with ϵ=ln(4) as they were at ϵ=ln(10), and another study showing people comfortable at ϵ=ln(7).

The challenges of differential privacy are real and if we can’t get epsilon to be a value that is usable, people (especially advertisers) may find other reasons to shy away from differential privacy or just ignore it. By **increasing epsilon while still respecting the user**, we increase the viability of differential privacy as a method and have a greater chance of better adoption of differential privacy which is the only way to get strong guarantees on individual user data.

Because any nonzero epsilon is infinitely better (literally!) than the alternative of no differential privacy.

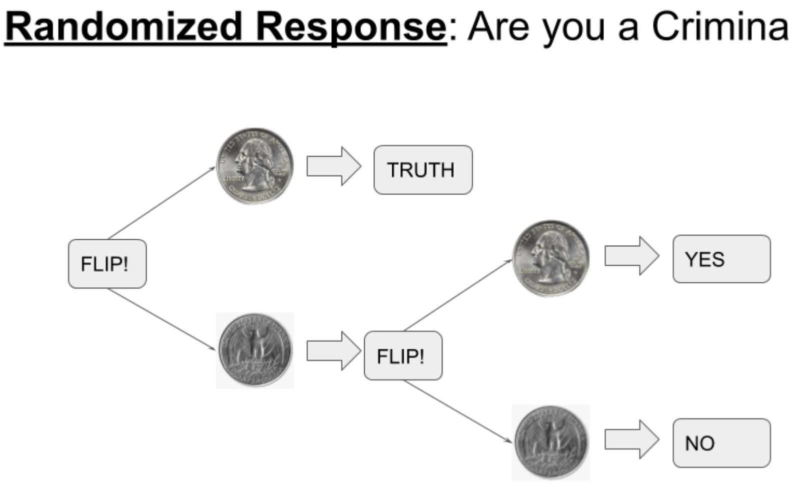
(Note: There are a lot of diagrams in this doc so it is a lot shorter than the page count may suggest!)

Background

What is randomized response?

If a pollster asks you an embarrassing or sensitive question you're less likely to respond truthfully because humans are social creatures and, generally speaking, want other humans to like us. This is even a problem in electronic surveys because humans get worried that their data could escape and be used against them. This is also precisely why tech companies cares about privacy of user data, because our users care about their privacy and we care about our users.

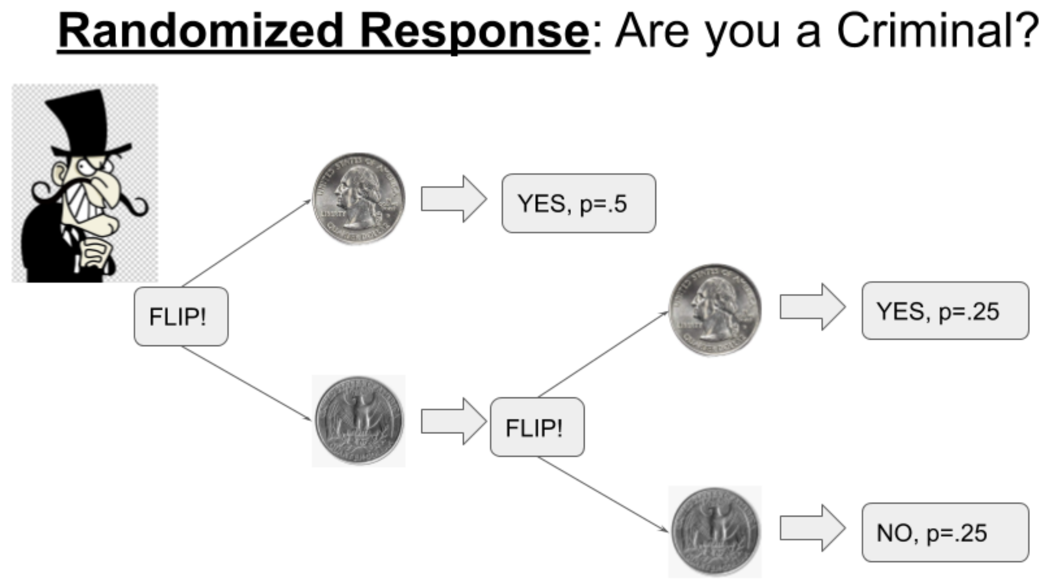
The solution is the "randomized response" technique, where you ask respondents to participate in random events to determine what their answer is. An archetypal scheme for randomized response looks like this:


Randomized Response: Are you a Criminal?

Because now anyone who answers true has plausible deniability that they, themselves, are not a criminal: they may have ended up flipping heads during the second coin flip instead of stealing a car in their twenties: the pollster truly can't know!

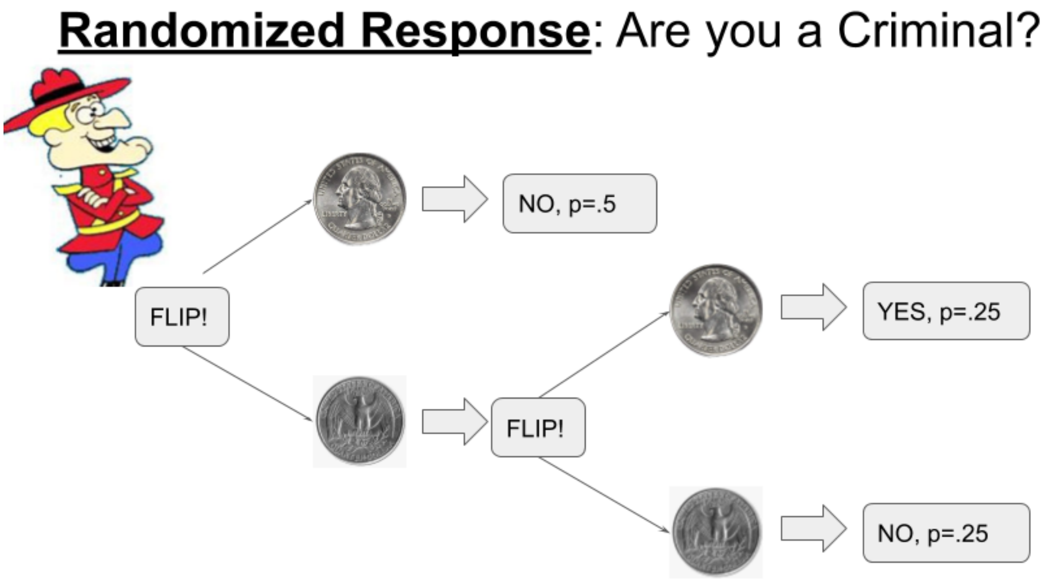
To now introduce the connection between differential privacy and randomized response. If you were to say that randomized response works, and people are OK with releasing truly embarrassing information to a pollster under these circumstances, then maybe if we look at this under the lens of differential privacy, then we'd get a good value for ϵ we'd feel comfortable protecting users with. And we can do just that!

Let's call the above picture our randomized mechanism M. Then, if you assume the neighboring databases are the user is a criminal and the user is a do-gooder, starting with the criminal case we'd have a flow chart like this:



Randomized Response: Are you a Criminal?

And the do-gooder case:



Randomized Response: Are you a Criminal?

And thankfully we only have 2 possible outputs of this random mechanism, so there’s just two ratios we have to consider:

$$r_1 = \frac{P[M(\text{criminal}) = \text{YES}]}{P[M(\text{not-criminal}) = \text{YES}]} = \frac{.75}{.25} = 3$$

$$r_2 = \frac{P[M(\text{criminal}) = \text{NO}]}{P[M(\text{not-criminal}) = \text{NO}]} = \frac{.25}{.75} = \frac{1}{3}$$

$$\epsilon = \ln(3)$$

And voila! That’s where one can get ϵ=ln(3) in a way from first-principles.

Is Randomized Response Great, though?

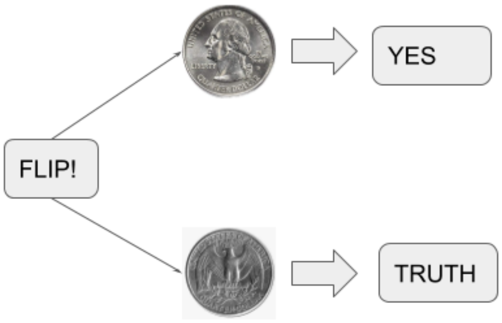
This should be great in theory, but the paper When and why randomized response techniques (fail to) elicit the truth (John 2018) calls into question the utility of the heretofore well accepted technique of randomized response in social science.

The paper basically concludes with skepticism that randomized response is any good for its intended purpose. And roughly speaking, this is because they have a well-accepted mechanism that fails to get higher prevalence estimates for sensitive questions than just asking participants the questions point-blank and they go through many well intentioned, rigorous improvements to their mechanism and every improvement fails to get a higher prevalence estimate than direct questions.

Why would you expect that? The whole point of randomized response is to get people more comfortable answering sensitive questions, instead of lying to make themselves look good. If it indeed does work, we would always expect randomized response to get a higher prevalence estimate than just asking people point-blank a sensitive question like “are you a criminal” because we expect more people to lie and say they’re not a criminal and avoid looking bad. The work in John 2018 is very interesting and they do seem to get better mechanisms as they iterated through their process. I’m not here, however, to pick that apart as it’s good work, but I have more of a meta-comment.

The problem, as I see it, with the paper is with the mechanism they were testing for all of their studies. If I were to draw a similar diagram to our coin flip above, it would look like this:

John 2018 Mechanism: Are you a Criminal?



And perhaps the reader can see this is not ϵ-differentially private! If the pollster gets a NO answer, they are 100% sure that the true answer is NO which is a privacy failure. In this case the motivation was good enough, they wanted more people to not have to transmit a bad-looking answer to the pollster. But in doing this, they created a way for their participants to with 100% certainty look good to the pollster. Perhaps the participants in their myriad studies could see this--most likely subconsciously. But the social human impulse to look good to other humans cuts both ways: they don't want to look bad but they also want to look good.

To analyze this precisely, let's look at this mechanism in terms of differential privacy. Depending on how you define **δ**, the failure probability, I would call this mechanism a **δ**=.5 algorithm. Then as to the question of ϵ:

$$r_1 = \frac{P[M(\text{criminal}) = \text{YES}]}{P[M(\text{not-criminal}) = \text{YES}]} = \frac{1}{.5} = 2$$

$$r_2 = \frac{P[M(\text{criminal}) = \text{NO}]}{P[M(\text{not-criminal}) = \text{NO}]} = \frac{0}{.5} = 0$$

$$\epsilon = \ln(2), \delta = .5$$

So to those of us in the privacy community, that value of ϵ is OK, but we'd never consider a value that high of **δ**.

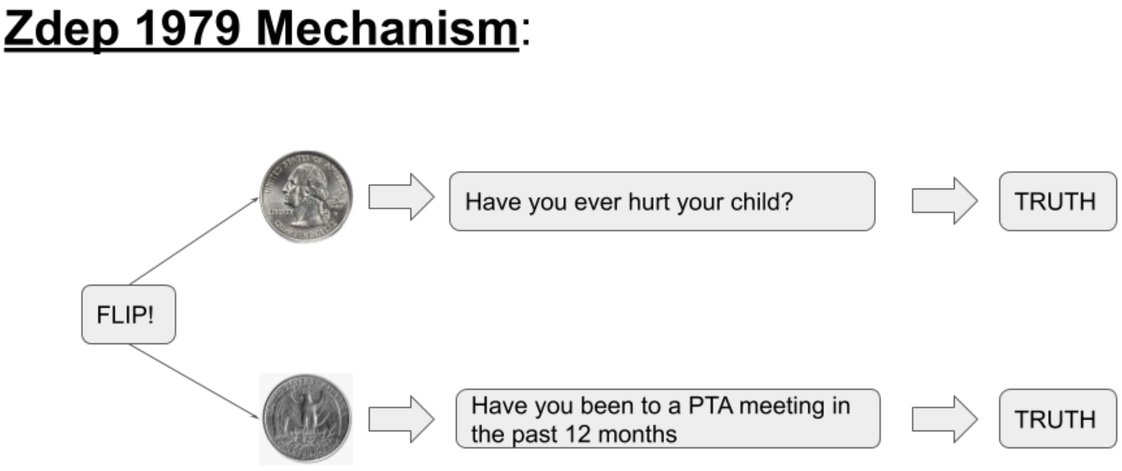
Hidden Confirmation in John 2018

(John 2018) went one further and actually published a list of the studies in their review which DID actually show higher prevalence estimates in their applications of randomized response (suggesting randomized response works) and even summarized the randomized response mechanisms for those studies. I wish more papers did this; it is truly outstanding scholarly citizenship: making it easier for people to look into why you might not be 100% correct. If you're following along in their paper, turn to appendix 2 to see this list.

To summarize: **all of the papers she listed as having found a favorable outcome for randomized response have a δ value of 0**. Other than their own studies, the authors of John 2018 did not similarly list the other studies they looked into that didn't show as good of an effect for randomized response, so this doc is not currently as ironclad as it could be if we also went into all the other cited mechanisms, but it's a start.

Zdep et al. (1979)

This is an interesting deviation from the randomized response explained up to this point, where instead of the response being randomly determined, the question is randomly determined. Like so:

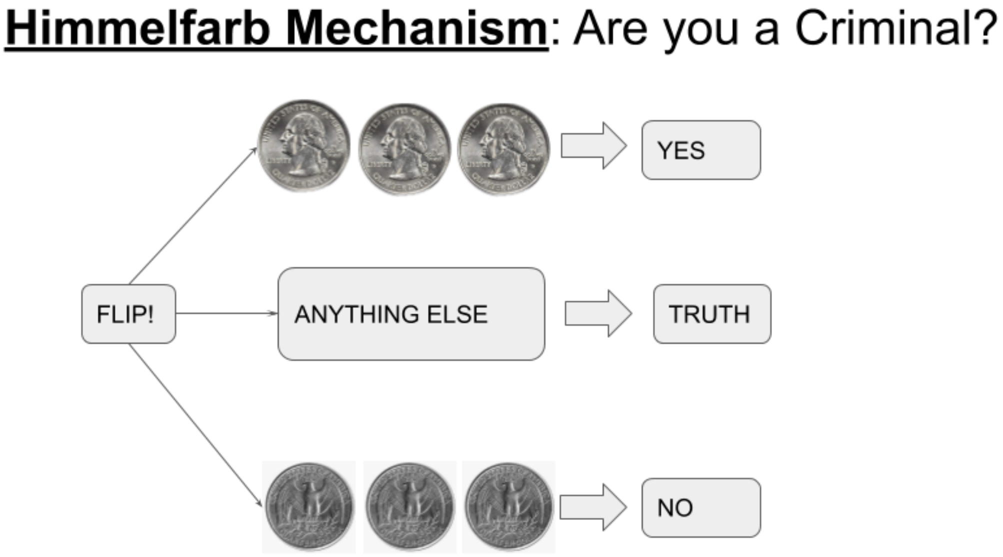

Zdep 1979 Mechanism:
FLIP!
Have you ever hurt your child?
TRUTH
Have you been to a PTA meeting in the past 12 months
TRUTH

(PTA means parent-teacher association: a common feature of schools in the United States)

I'm not sure it would be informative to try and rigorously formulate this like differential privacy although I am pretty sure that the ϵ here would look like ln(2). Please feel free to disagree in the comments!

Himmelfarb and Lickteig (1982)

This is a very interesting one in my opinion. They used 3 coins and if they ALL came up heads, or tails then you automatically give a specified answer, like so:



Himmelfarb Mechanism: Are you a Criminal?

FLIP!

YES

ANYTHING ELSE

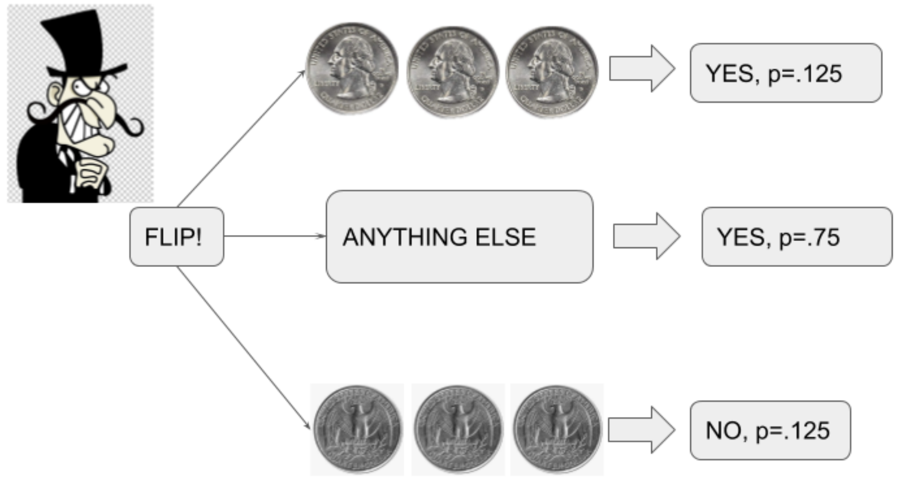
TRUTH

NO

Which with three coin flips is unique to everything we've considered before with only 1 or two coin flips.

We can then look at the criminal and do-gooder cases:

Himmelfarb Mechanism: Criminal Case



FLIP!

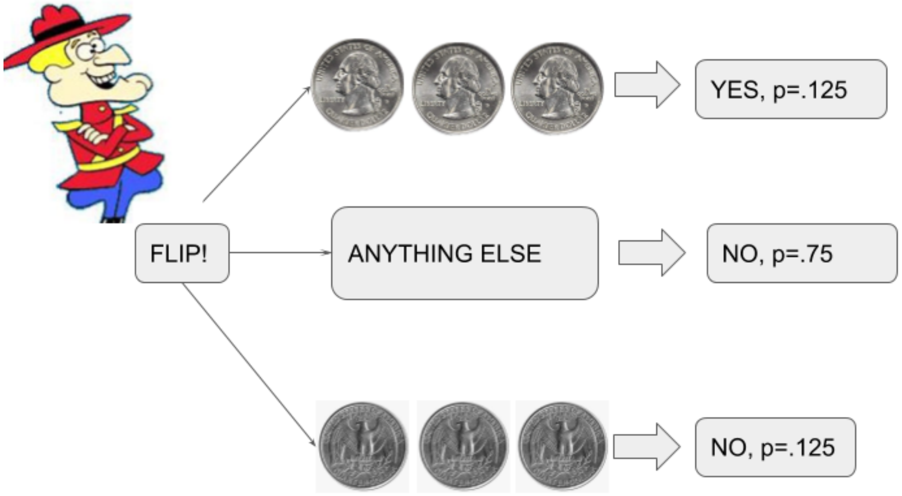
YES, p=.125

ANYTHING ELSE

YES, p=.75

NO, p=.125

Himmelfarb Mechanism: Do-Gooder Case



FLIP!

YES, p=.125

ANYTHING ELSE

NO, p=.75

NO, p=.125

Then doing the math for this mechanism:

$$r_1 = \frac{P[M(\text{criminal}) = \text{YES}]}{P[M(\text{not-criminal}) = \text{YES}]} = \frac{.75 + .125}{.125} = 7$$

$$r_2 = \frac{P[M(\text{criminal}) = \text{NO}]}{P[M(\text{not-criminal}) = \text{NO}]} = \frac{.125}{.75 + .125} = \frac{1}{7}$$

$$\epsilon = \ln(7)$$

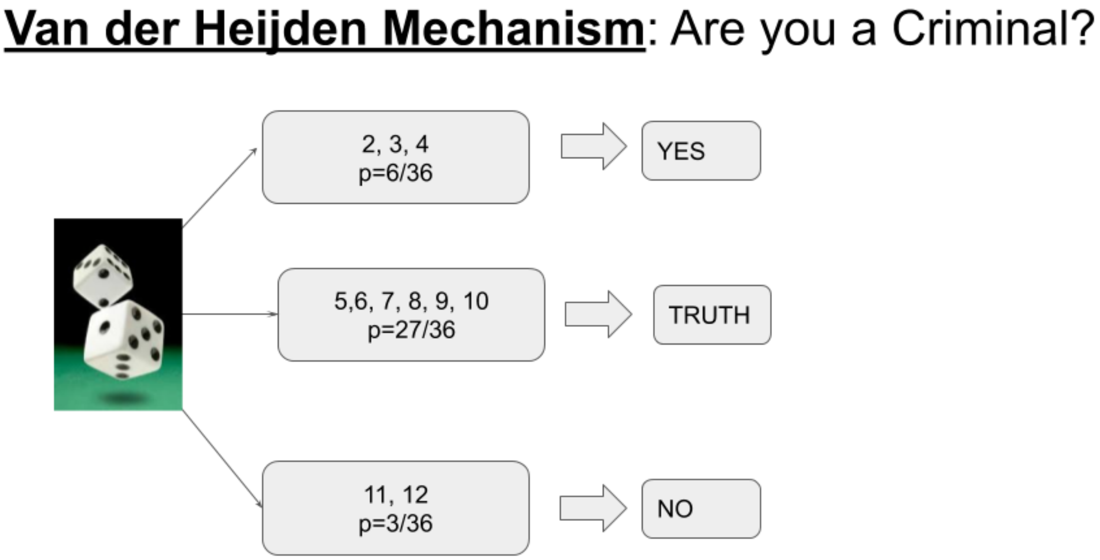
Note that the paper, given its age, is not available for free via google scholar so I have not yet done a deep dive on this paper. But if they truly did get good results from people on sensitive questions with an epsilon of ln(7).

Barth and Sandler (1976)

Barth and Sandler is a mix of both Himmelfarb with 2 coins, and Zdep because they have the subjects flip two coins, and if they both come up heads they answer an innocuous question "does your telephone number end in an odd digit", and any other combination, they answer if "in the past year, have you consumed 50 or more alcoholic drinks". Much like with Zdep, I will not go in depth of the math for this mechanism, but it's likely a delta=0 mechanism at least if you assume the researchers don't know the phone numbers of the participants.

Van der Heijden et al. (2000)

This one gets even more interesting, because they used two dice! Which in the style of our earlier experiments looks like this:



Van der Heijden Mechanism: Are you a Criminal?

2, 3, 4
p=6/36 → YES

5,6, 7, 8, 9, 10
p=27/36 → TRUTH

11, 12
p=3/36 → NO

It's a little lopsided, I'm not sure why. If I had to guess it was likely to improve the error of the estimator. But regardless, the differential privacy math would look like so:

$$r_1 = \frac{P[M(\text{criminal}) = \text{YES}]}{P[M(\text{not-criminal}) = \text{YES}]} = \frac{33/36}{6/36} = 5.5$$
$$r_2 = \frac{P[M(\text{criminal}) = \text{NO}]}{P[M(\text{not-criminal}) = \text{NO}]} = \frac{3/36}{30/36} = \frac{1}{10}$$
$$\epsilon = \ln(10)$$

So compliance with this experiment would suggest that ln(10) is a reasonable protection for user data! It is the last paper listed by John 2018 and the highest value of ϵ considered so it's worth diving a bit deeper into, especially since I could find the paper for free on google scholar!

One Study: two epsilons, ~same compliance!

As mentioned above, I'm going to deep dive into Van der Heijden et al. (2000). And spoiler alert, two of their studies had different epsilons, and they had roughly the same prevalence estimates (indicating the same level of comfort releasing data) but one had epsilon=ln(4), the other had epsilon=ln(10).

You can find a pdf of the paper here to follow along. It's a very well written paper, I highly encourage you to read it.

They were asking participants questions about welfare, which the researchers argue is a sensitive topic and in particular they were trying to learn ways to reduce fraud in the welfare system. This paper also has a good survey of randomized response.

The survey was ostensibly just about how people on welfare get by, but they had fraud questions mixed in. Further adding to the study, their participants had all been caught by the dutch government for welfare or unemployment benefit fraud, but the interviewers did not know that their subjects had that distinction, and the subjects didn't know they were set up in this way either.

The researchers also crucially noted that many of these people may very well have only failed to meet a deadline for telling the welfare office that they are newly employed and thus are only guilty of missing a bureaucratic deadline. So the participants might not consider themselves to have done anything wrong so the true prevalence is not fully 100%.

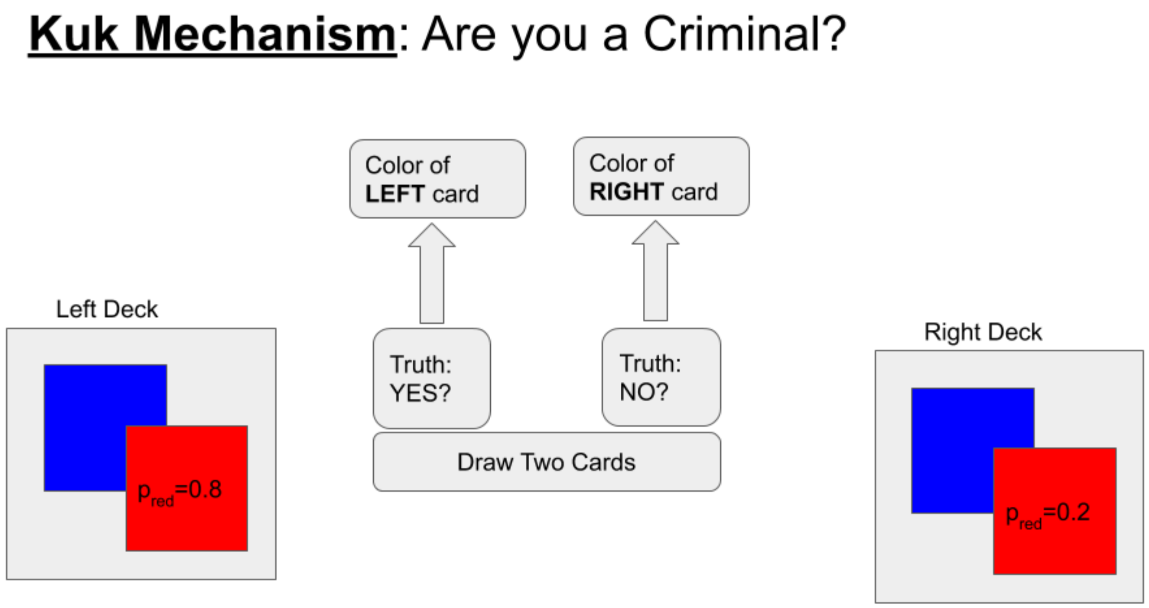
They also picked some subjects with a below average education level and a lot of non-native dutch speakers.

Anyone who had "extreme difficulty" doing the randomized response mechanisms, they did put into the face-to-face direct questioning bucket. But given they have a rough idea for the underlying truth for everyone I suspect that shouldn't bring the study's conclusions into question too much.

The response rate was close to 50% from all the people they initially approached, and you could argue that people who are even more private would have been more likely to refuse to participate, thus skewing the study in the direction of randomized response working better than it should. But I personally would say the more likely explanation is laziness for not responding to the researchers. You might also argue that the dutch would have different opinions on privacy than others and that may be true, but they included lots of immigrants in the study so that would be less clear to me also.

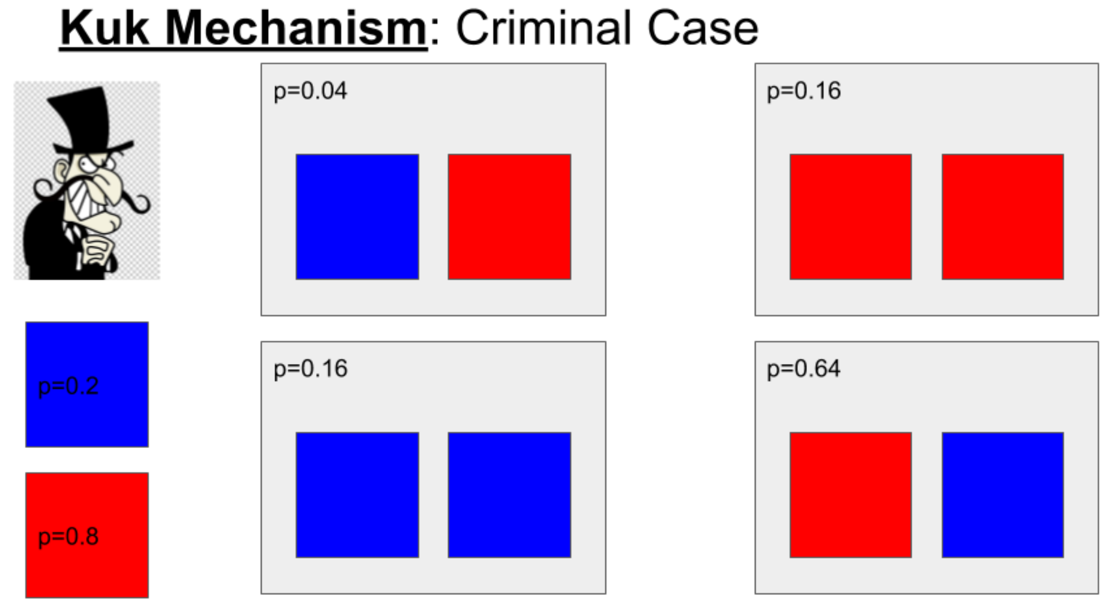
Kuk Mechanism

In addition to the above-calculated randomized response method, they also used another method to compare, called Kuk's procedure which is really cool and yet another application of randomized response. They used this because the other technique they used, a criticism of it still requires the user to "tell the truth" in some cases, but this one adds a very interesting additional layer of complexity. I've drawn the diagram out here:



You have two, unbalanced decks of red and blue cards. You then have the subject draw two cards, and give the color of the left card if their answer is yes, or the right card if their answer is no.

This is a bit more complex so I'll bring back explicit diagrams to calculate the ϵ guarantee of this mechanism. First for a criminal:



Kuk Mechanism: Criminal Case

🟦 p=0.2

🟥 p=0.8

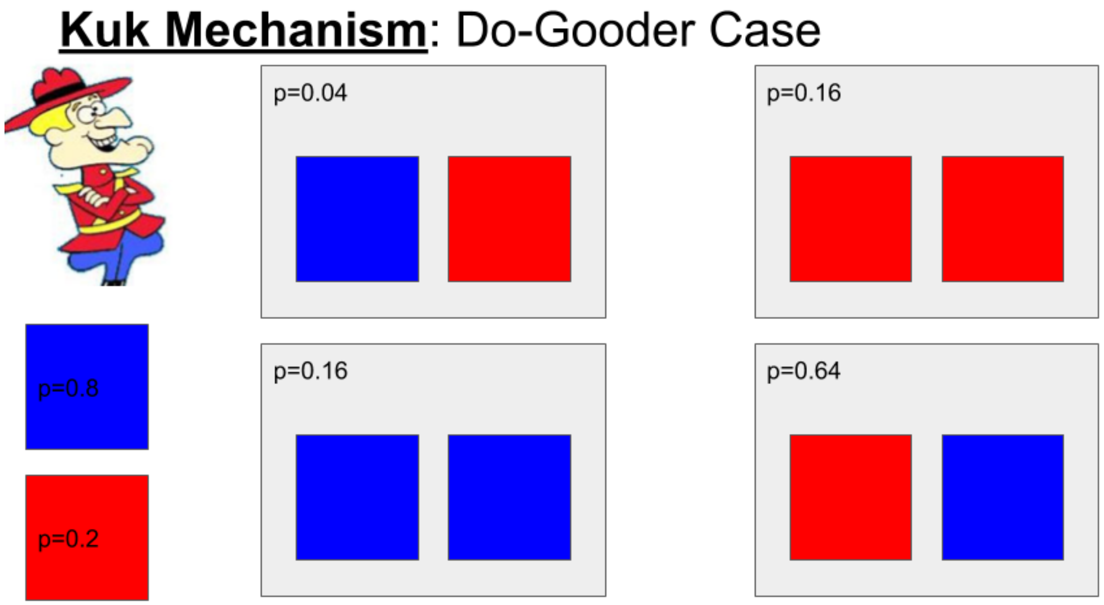
p=0.04 🟦 🟥

p=0.16 🟥 🟥

p=0.16 🟦 🟦

p=0.64 🟥 🟦

Then for a do-gooder



Kuk Mechanism: Do-Gooder Case

🟦 p=0.8

🟥 p=0.2

p=0.04 🟦 🟥

p=0.16 🟥 🟥

p=0.16 🟦 🟦

p=0.64 🟥 🟦

And then the math is then:

$$r_1 = \frac{P[M(\text{criminal}) = \text{Red}]}{P[M(\text{not-criminal}) = \text{Red}]} = \frac{.8}{.2} = 4$$

$$r_2 = \frac{P[M(\text{criminal}) = \text{Blue}]}{P[M(\text{not-criminal}) = \text{Blue}]} = \frac{.2}{.8} = \frac{1}{4}$$

$$\epsilon = \ln(4)$$

It’s interesting that Van der Heijden actually noted in their results section that they actually got a slightly higher prevalence estimate than the aforementioned response technique which is commonly called “forced response” and had a higher ϵ. I think it’s not necessarily fair to interpret the difference in response to the difference in ϵ especially because it’s such a small difference to likely be noise level.

I’m going to grab the most relevant data table front he paper and reproduce here:

Method	epsilon	n	P (fraud)	SE	z
Main Mechanism	ln(10)	96	.43	.068	2.22
Kuk Mechanism	ln(4)	105	.49	0.82	2.59
Computer-assisted self interview	infinity	47	.058	.058	-0.83
Face-to-Face	infinity	99	.25	.044	

So while that is a 12% lower fraud estimate isn’t nothing, they are very close indeed. Furthermore, given the quoted standard errors, the fact these are random mechanisms to start with, and the closeness of z-test statistics relative to the face-to-face answer, it’s reasonable to guess that they might not be statistically different from each other.

Likely Objections

“But wait! Have you thought about ___?”, you might be thinking. I likely have been thinking about it too, so here are some discussions on weak parts in my arguments above and what I’ve thought about them. Please contribute to the discussion, though, and point out anything extra you might be able to think of!

Neighboring Database Definition

Normally, we consider the unit of protection a deletion of the user from a dataset. But in all of the analyses up to this point, we treat what looks more like a switching of a user out to a different user, which is equivalent to a deletion+addition. This is a problem for my argument because naively that means that if, epsilon_deletion=epsilon_swap/2 and ln(10)/2 is only 5%

higher than ln(3) so useful but not the twice as useful as we might have been hoping for. So have I just argued only for a 5% boost in epsilon? I think not and let me explain why.

Differential privacy relies on the assumption that a central authority holds the data and is doing data releases that respect user privacy. In randomized response, however, the definitions are flipped a bit around. In Randomized response, the database holder is the user and the researcher is querying the database no matter what. The researcher is then instructing the subject/database-holder to release data in a method that the researcher can't tell with absolute certainty the neighboring universes of the subject being a criminal or not a criminal. So the more appropriate dataset definitions are "the subject is in the universe of criminals" and "the subject is NOT in the universe of criminals". Because they've already elected to participate, the subject/user then subconsciously knows that they are only releasing epsilon of information related to whether they are in the sensitive group of subject/users

But that's exactly what the data owner is protecting! Our preferred neighboring datasets boils down to things like "this user is in the universe of people who saw this coke ad" versus "this user is NOT in the universe of people who saw this Coke Ad". So here, we can assume our users are telling us they are OK with us doing data releases the same way they were ok with doing data releases: such that the people we are telling can't tell if the user is a member of the set of users who saw this coke ad, versus a member of the users who didn't see this coke ad.

To summarize, I'll put my argument into table form

	Randomized Response	**Central Data Release**
Database Holder	User	Data Owner
Sensitive dataset	User $\in$ {criminals}	User $\in$ {Saw Coke Ad}
Non-sensitive Dataset	User $\in$ {non-criminals}	User $\in$ {No Coke Ad}
Relevant Epsilon	ϵ	ϵ

(the last row mainly signifying that the epsilon of a data release should be the same as the epsilon of a user data release to a researcher in randomized response, not half of that epsilon)

People Don't Understand Probability let alone Differential Privacy

I only ever mean to be arguing that people SUB-conciously can tell that their data is being protected to a lesser or greater extent. And people could likely tell in orders of magnitude the amount of uncertainty. If you told participants to instead roll 1 million dice and if they all come up 1, then say yes, if they all come up 6 say no, otherwise tell the truth… I'm guessing they would get that is basically the same as telling the truth.

Data Authority != Dutch University Researchers

The truth is that in the limit that everyone understands differential privacy, every human being will have their own comfortable epsilon for every different data provider for every different kind of data. One can imagine that culture at the country level might be the principle component of an individual's privacy preferences, but everyone is still an individual.

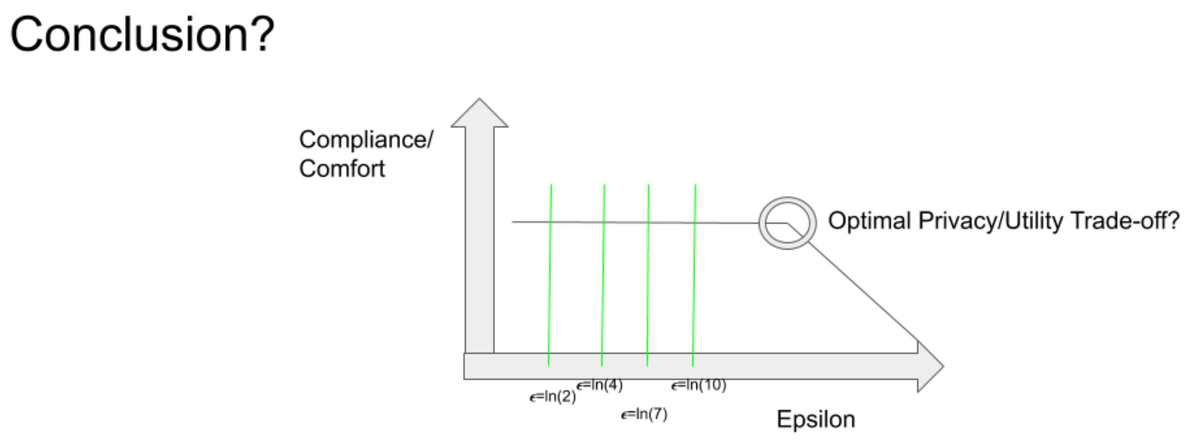
And really the strongest piece of evidence presented in this doc is the relative constant compliance between ln(4) and ln(10) in the exact same study on the exact same sensitive information.

Summary/Conclusions

I think John 2018 did a pretty good job showing that the randomized response which has a large $\boldsymbol{\delta}$ does not have a significant benefit above just asking directly sensitive questions. And comparing their work with $\boldsymbol{\delta}$=.5 with other work they cited where $\boldsymbol{\delta}$=0, we do see a more people willing to answer truthfully the answers to sensitive questions.

Given also as recently as 2018, the randomized response social science research community was not aware of differential privacy, that there could be a highly fruitful collaboration between those two communities.

I put it that the papers we have considered, basically builds a graph like this:



Conclusion?

Van der Heijden then shows pretty convincingly that people will also generally be honest with an ϵ of ln(10) and basically the same level of honest as they are at what is approximately equal to ln(3).

But those researchers are also asking about a crime that was committed and they knew it would be published verbatim to the general public. So that data is generally more sensitive than data about online activity, but probably generally less sensitive than things like location data which could be used to commit a crime against you.

I am just saying that we have evidence people are fine with epsilon=ln(10) so we instead of having an arbitrary and vague sense we want to protect users to a 50% chance of being right, we have hard evidence that people were fine with an epsilon=ln(10) mechanism and thus we have a completely non-arbitrary number with a great justification and further that humans participating in the study were at least subconsciously as OK with ln(4) as they were with ln(10).

Furthermore **the challenges of differential privacy are real and if we can't get epsilon to be a value that is usable, people will find other reasons to shy away from differential privacy or just ignore it. By increasing epsilon while still respecting the user, we increase the viability of differential privacy as a method and have a greater chance of better adoption of differential privacy and protecting more user data by the best way yet discovered.**