Estimating the Number of Bots on Twitter using Idena Protocol

Andrew Edi, co-founder of Idena Proof-of-Person blockchain
info@idena.io

Abstract. The paper proposes a solution for the estimation of the number of bots and fake accounts on Twitter using the open source Idena Proof-of-Person protocol. Idena protocol employs a synchronous ceremony for simultaneous verification of unique users with AI-resistant flip-tests.

Introduction

The proposed solution is to give the estimation of the number of real users interacting with the Twitter app excluding the accounts driven by the bots (algorithms) and click-farms' accounts, where many Twitter accounts are managed by a single person using many devices at the same time. The method will not lead to an exact number of bots on Twitter, but gives an estimate on the order of magnitude (e.g. is it 5% or 25%).

The solution does not require the Twitter users personal data (id documents, videos, photos, etc.). The solution is not affected by the language the user speaks.

Proposed solution

We propose to run a series of trials T for the randomly selected set of users at different times of the day and on different days of the week. For the single trial $t \in T$, a subset of Twitter accounts u_t is randomly selected from the whole set of users U_t that are actively interacting with the Twitter app (for example, scrolling through the feed) at the certain time when the trial t is started.

For a selected subset u_t of challenged users, a Twitter app runs a Turing test in the form of AI-resistant flip-tests (see details below). For the single trial t, all the selected users u_t are simultaneously challenged. This eliminates the possibility of false-positive verification of the multiple accounts managed by a single person (to exclude click-farms' accounts).

For one single trial t, we will get the ratio R_t of successfully verified accounts. For a series of trials T, we will get the average ratio R_T of successfully verified users, which in the end will reflect the statistical estimation of the unique live Twitter users.

Adjusted estimation

Average ratio R_T estimates the number of Twitter accounts that successfully passed the synchronous tests. However if some of the users declines the test it leads to an incorrect result. We propose to calculate an adjusted estimation R of the number of real users. For that we conduct asynchronous tests among a reference group of verified Twitter users V, for whom it's determined that they are not bots and their accounts are not managed by click farms with a high probability.

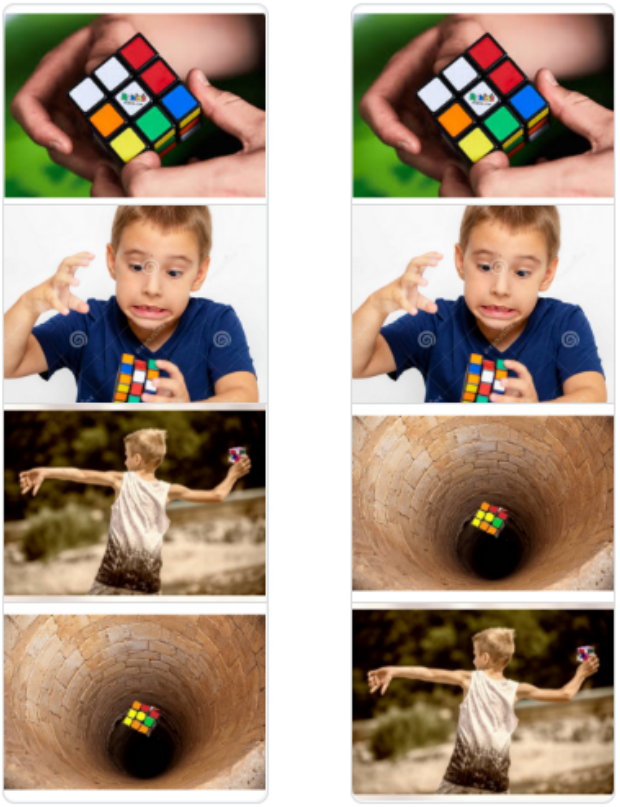
As a result of testing of users in the reference group V, we will get a success ratio R_V for the verified Twitter users. Adjusted estimation R is calculated as follows:

$$R = \frac{R_T}{R_V}$$

Idena Flip-test

FLIP stands for the "Filter for Live Intelligent People". Flip is not an IQ test but a common sense test. A flip is a language-neutral AI-resistant kind of CAPTCHA.

Flip utilizes four images. To solve a flip, the user chooses between two sequences of these images, only one of which makes narrative sense. The other one is deliberately distorted so that the picture sequence does not convey linear story information.



A meaningful story (left) and a meaningless sequence of pictures (right)

To pass a test challenged users must successfully solve at least 5 out of 6 flips. They have 120 seconds to complete the test. Before starting the time-constrained test we recommend showing a countdown timer and a motivational call to take the test to help us determine the number of bots on Twitter.

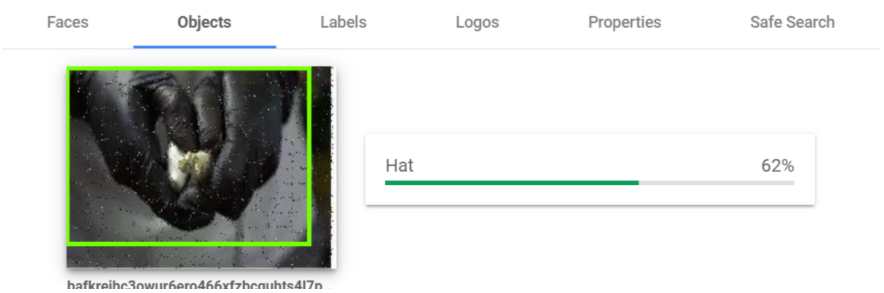
AI-resistance of the flips

Instead of the widely used Google reCAPTCHA or other types of CAPTCHAs based on proprietary algorithms we propose the use of unique Idena flips which are handcrafted by a decentralized network of people. This makes it difficult for bots to solve them. The Idena flips are obscure on the web. There are around 1M of flips on the Idena public blockchain which are encrypted and can not be used for AI training.

Adversarial attacks can be applied to the flips to prevent solving them by an AI. Since the flip is a common sense test there are two layers of adversarial attacks possible:

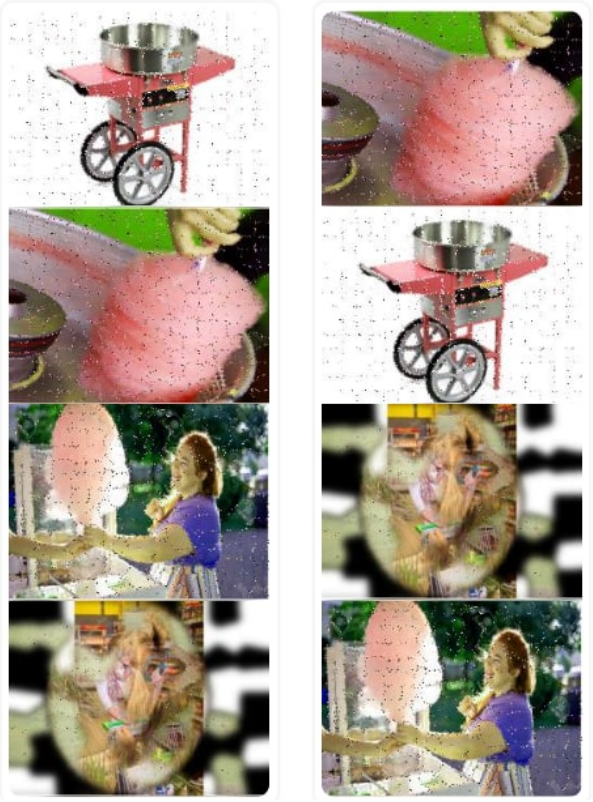
1. **Adversarial perturbations** are added for each of the 4 images of a flip to make it harder for AI to classify the images.

Adversarial perturbations applied to the images of the flip



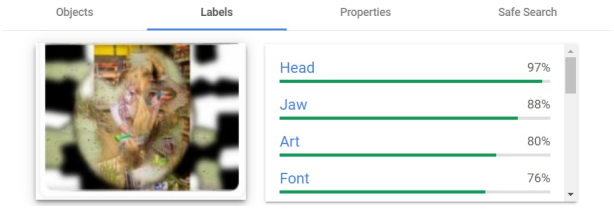
Google Vision result for the image with adversarial perturbation: Hat while original image is classified as Gloves

2. **Adversarial nonsense images** are used to make it harder for AI to determine which sequence of images makes sense.



A flip with an adversarial nonsense image

Nonsense images added into flips do not stop people from solving them. In contrast, it makes it harder for AI to solve the flips.



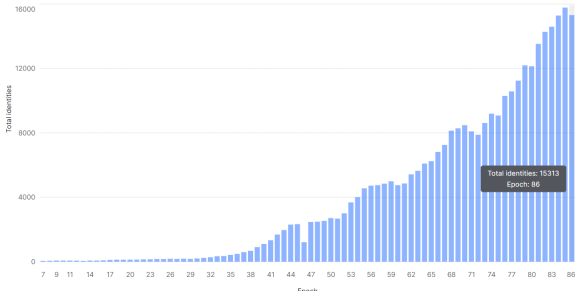
Nonsense images are classified as meaningful

Meaningful classification of the nonsense image leads to an unpredictable outcome for an AI that solves the entire flip based on the wrong image classification data.

About Idena

Idena is the democratic network of people based on the Proof-of-Person blockchain. Every Idena node is linked to a cryptoidentity – one single person with equal voting power and mining income.

Any human can become an Idena validator by proving their uniqueness and humanness during the Idena validation ceremony. The ceremony runs once per 3 weeks on Saturdays at 13:30 UTC. There are around 13k participants and the number is growing.



Total number of validated identities in Idena network

The Idena blockchain is secured by the Proof-of-Person consensus. There are no centralized servers. The network consists of 1.5k independent validating nodes. To scale the possible number of participants and the blockchain performance the network sharding is used.

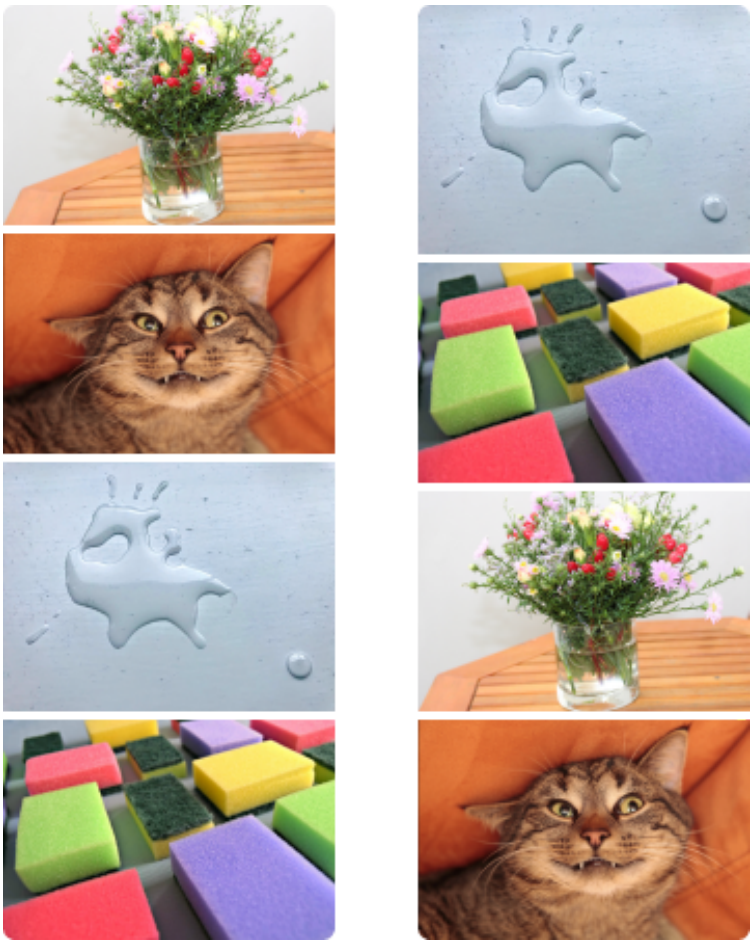
There are around 1M flips created on the Idena blockchain. However most of them are encrypted. To run the proposed test on Twitter the Idena core team can provide a collection of 15K unencrypted flips. Potentially the Idena network could be used to generate new flips designed especially for the test on Twitter.

For each flip test, there is a consensus of 4-12 participants representing the correct answer and the probability that a person can successfully solve the flip.

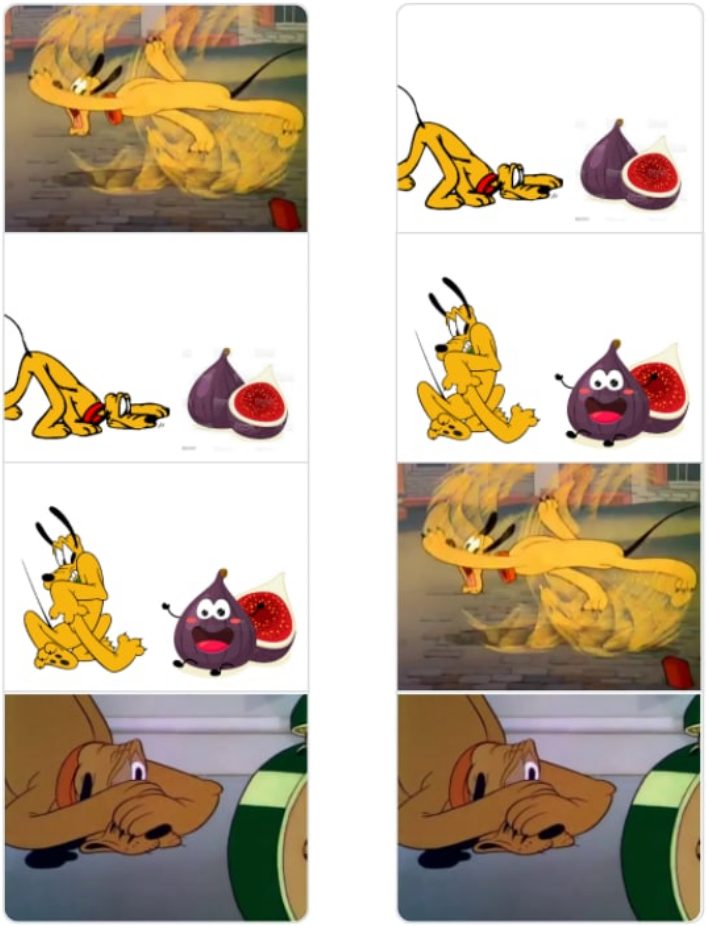
References

1. Idena web-site: https://idena.io
2. Idena blockchain explorer: https://scan.idena.io/
3. Idena White Paper: https://docs.idena.io/docs/wp/summary/

Appendix: flip examples

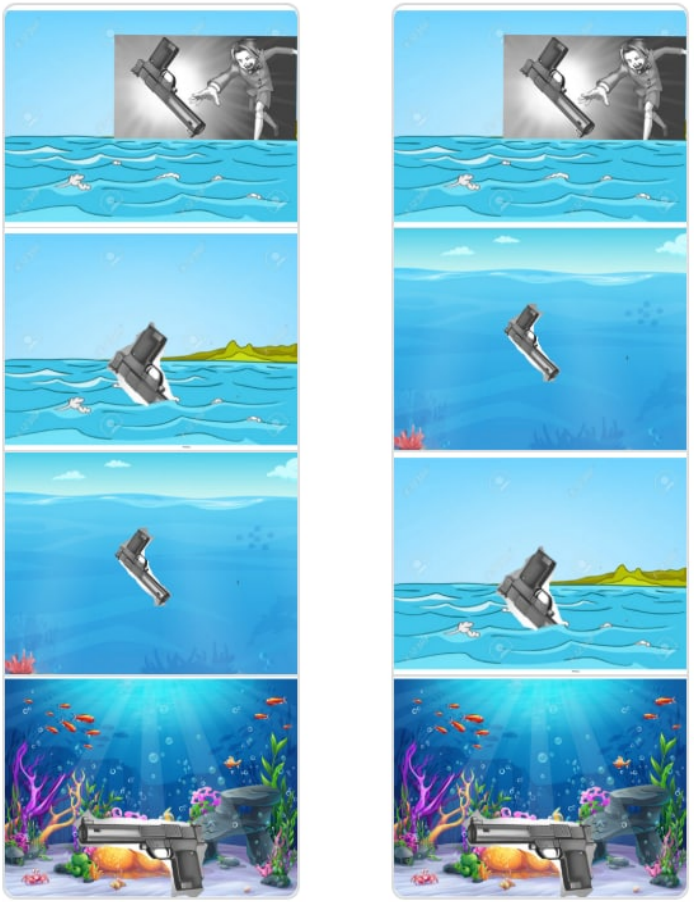


Meaningful story: left

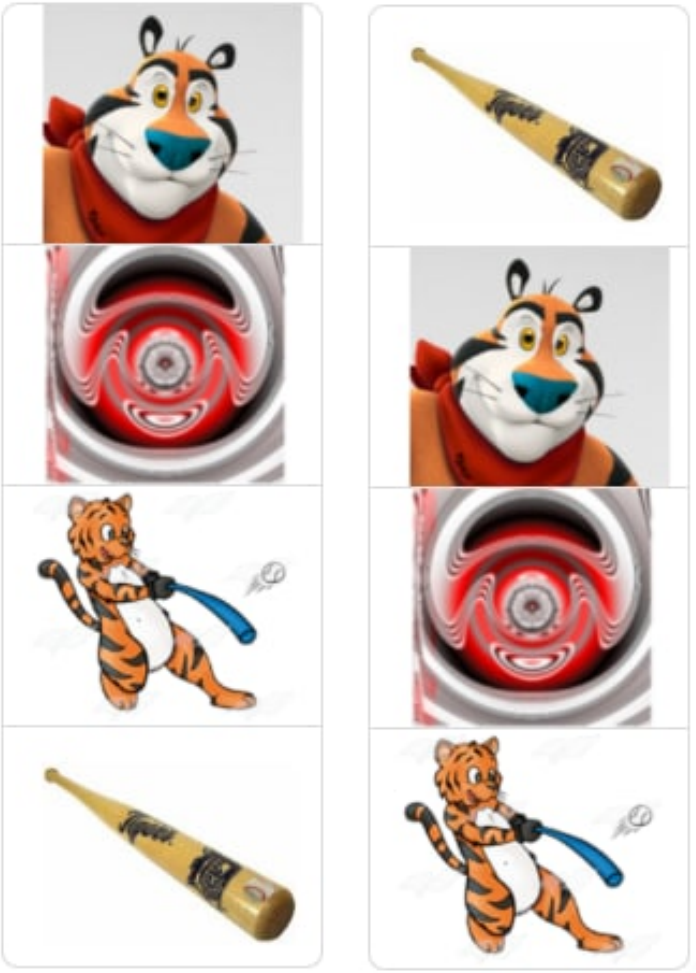


Meaningful story: right

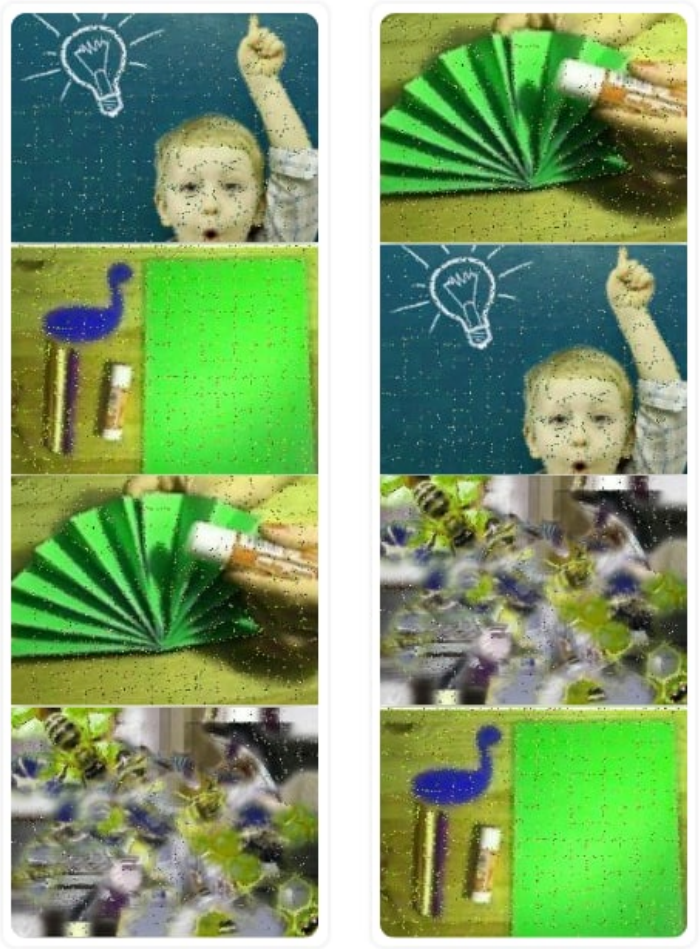
Meaningful story: left



Meaningful story: left



Flip with an adversarial nonsense image. Meaningful story: right



Flip with adversarial perturbations and the nonsense image. Meaningful story: left