

**Supporting Information:
The Shape of and Solutions to the MTurk Quality Crisis**

August 7, 2019

Table of Contents

Page 2. A1: Model Details.

Page 6. A2: Comparison of data quality under different screening conditions

Page 8. A3: R package for interaction with IP Hub

Page 12. A4: Qualtrics screening protocol for blocking VPS and international respondents

Page 23. A5: A note on human subjects compliance

Page 24. A6: An analysis of individual quality checks

Page 28. A7: Analysis of ISP and timing of survey responses

A1 Model Details

Table A1, below, shows the results of OLS models predicting political knowledge scores and the logged time spent on the political knowledge questions in retrospective study 1.

Table A1. Political knowledge as a function of IP status

| | Knowledge Score | | Time on Knowledge | |
|--------------------|-----------------|----------|-------------------|----------|
| VPS User | -0.28 * | -0.40 ** | 0.97 *** | 0.74 *** |
| | (.14) | (.14) | (.10) | (.10) |
| VPS Uncertain | 0.28 | 0.30 | 0.33 | 0.34 |
| | (.37) | (.36) | (.26) | (.24) |
| Foreign IP | 0.04 | 0.02 | 0.78 *** | 0.55 *** |
| | (.22) | (.22) | (.16) | (.15) |
| Political interest | - | 0.27 *** | - | -0.03 |
| | | (.05) | | (.03) |
| Education | - | 0.03 | - | 0.03 |
| | | (.04) | | (.02) |
| Male | - | 0.17 | - | -0.07 |
| | | (.09) | | (.06) |
| White | - | -0.08 | - | 0.05 |
| | | (.10) | | (.07) |
| Time on survey | - | - | - | 0.56 *** |
| | | | | (.05) |
| Constant | 2.72 *** | 1.60 *** | 3.17 *** | -0.16 |
| | (.05) | (.23) | (.04) | (.36) |
| N | 576 | 575 | 576 | 575 |
| R-squared | 0.01 | 0.08 | 0.16 | 0.31 |

Table A2, below, shows the results of OLS models predicting partisan identification as a function of political ideology, IP status, and interactions between each IP status and ideology. Valid respondents represent the omitted status condition.

Table A2. The relationship between ideology and partisan identification by IP status

| | Partisan Identification | |
|--------------------------|-------------------------|---------------------|
| | Study 1 | Study 2 |
| VPS User | 2.46 *** (0.40) | 2.97 *** (0.30) |
| VPS Uncertain | -0.12 (0.96) | -1.28 (1.31) |
| Foreign IP | 1.37 * (0.60) | 1.60 *** (0.41) |
| Ideology | 1.00 *** (0.03) | 1.34 *** (0.04) |
| Ideology x VPS User | -0.45 *** (0.08) | -0.82 *** (0.08) |
| Ideology x VPS Uncertain | -0.04 (0.22) | 0.43 (0.39) |
| Ideology x Foreign IP | -0.46 ** (0.16) | -0.62 *** (0.13) |
| Constant | -0.08 (0.13) | -0.14 (0.11) |
| N | 576 | 1180 |
| R-squared | 0.64 | 0.57 |

Table A3. Experimental results by respondent IP status.

| | Full Sample | Valid Sample | Fraudulent Sample |
|-------------|--------------------|--------------------|----------------------|
| Treatment 1 | 0.75 ** (.07) | 0.79 *** (.07) | 0.55 *** (.17) |
| Treatment 2 | 0.63 *** (.06) | 0.70 *** (.07) | 0.43 ** (.15) |
| Treatment 3 | 0.28 *** (.06) | 0.27 *** (.06) | 0.35 * (.16) |
| Treatment 4 | 0.22 *** (.06) | 0.22 *** (.06) | 0.21 (.14) |
| Treatment 5 | 0.11 (.06) | 0.04 (.06) | 0.49 ** (.15) |
| Treatment 6 | 0.03 (.06) | -0.01 (.06) | 0.26 + (.15) |
| Treatment 7 | -0.11 * (.05) | -0.16 ** (.06) | 0.15 (.15) |
| Treatment 8 | -0.50 *** (.06) | -0.60 *** (.06) | -0.01 (.17) |
| Treatment 9 | -0.41 *** (.06) | -0.54 *** (.06) | 0.25 (.18) |
| Vignette 2 | -0.03 (.04) | -0.02 (.04) | 0.01 (.12) |
| Vignette 3 | 0.04 (.04) | 0.04 (.05) | 0.06 (.11) |
| Vignette 4 | -0.08 | -0.09 * | 0.06 |

| | | | |
|--------------|----------|----------|----------|
| | (.04) | (.04) | (.11) |
| Vignette 5 | 0.11 ** | 0.10 * | 0.21 |
| | (.04) | (.04) | (.12) |
| Vignette 6 | -0.13 ** | -0.15 ** | -0.01 |
| | (.04) | (.05) | (.12) |
| Constant | -0.06 | 0.00 | -0.37 ** |
| | (.04) | (.05) | (.13) |
| Respondents | 576 | 480 | 96 |
| Observations | 3456 | 2880 | 576 |
| R-squared | 0.23 | 0.29 | 0.06 |

A2 Comparison of data quality under different screening conditions

As was noted in the main paper, there were some differences between those who are labeled as potentially problematic by IP Hub, AbuseIPDB, and IPVOID. From this, the question naturally arises of how these differences affect the quality of the data that passes through these three screeners?

To see how IP Hub compares with these other services, we ran the first retrospective study through both IP Hub and the two services linked through Know Your IP (27). The results are in Table A4. As is clear, IP Hub produces similar results to AbuseIPDB, which is designed to track similar profiles. They agree on 97.3% of the cases. Conversely, IPVOID does not correspond with the results from IP Hub very well, agreeing on only 88.5% of cases. But, as noted in the main paper, this is likely because IPVOID's blacklist is not directly aimed towards VPS detection.

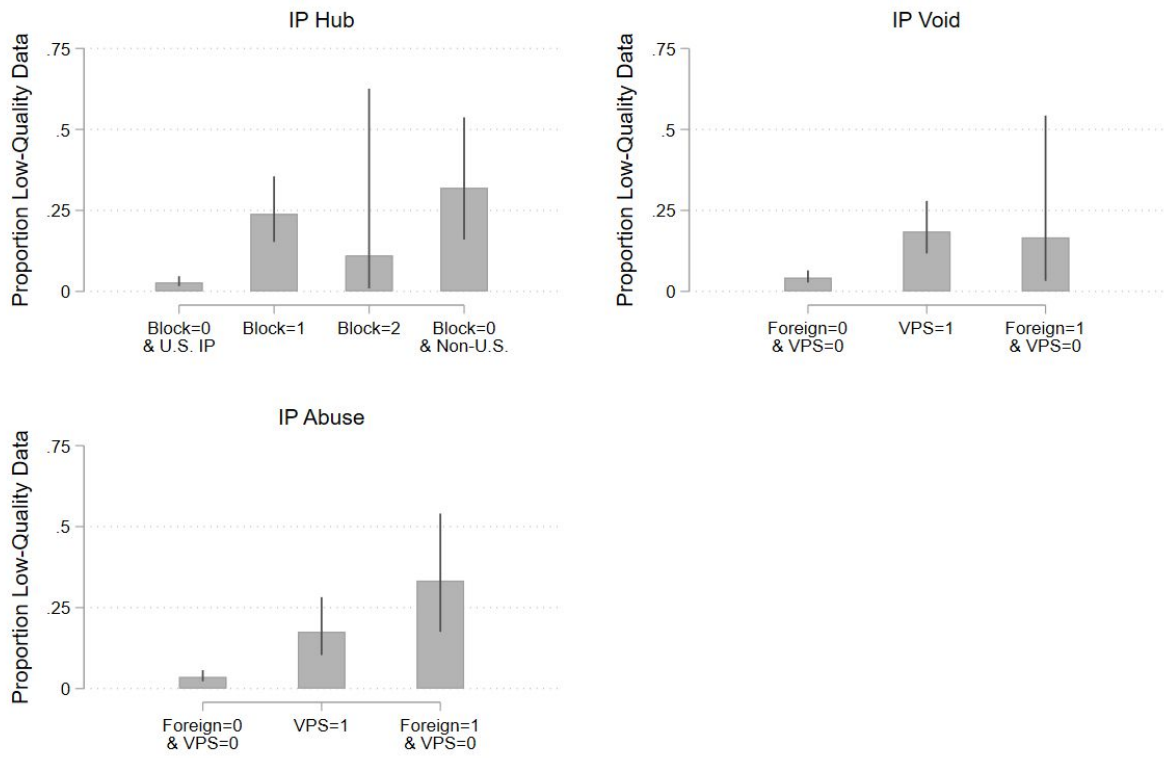
Table A4: Comparison Between IP Hub and Know Your IP

| | AbuseIPDB Block | AbuseIPDB Safe | IPVOID Block | IPVOID Safe |
|--------------|--------------------|-------------------|-----------------|----------------|
| IP Hub Block | 91 | 5 | 67 | 29 |
| IP Hub Safe | 10 | 470 | 37 | 443 |

Nevertheless, as we show in Figure A1, those labeled as clean in all three datasets have approximately the same performance on our quality checks. As described in the manuscript, we included five data quality checks, such as whether respondents consistently reported their age and location and whether they could describe their task in the survey. We focus on a dichotomous indicator of whether a respondent was flagged as providing low-quality data according to at least one of the five measures. Following the analysis in the manuscript, we analyzed data quality among different IP types using each of the three alternative classifications schemes (IP Hub, AbuseIPDB, IPVOID).

The results are shown in Figure A1. As is clear, the three different IP services provide largely similar results. Respondents who were not flagged as fraudulent provided high-quality data across all three services (IP Hub: 2.8%, AbuseIPDB: 3.6%, IPVOID: 4.2%). All three services show higher rates of low-quality respondents among those who were flagged as outside of the U.S. (IP Hub: 32.0%, AbuseIPDB: 33.3%, IPVOID: 16.7%). All three services also show higher rates of low-quality respondents among those who are flagged for VPS use (IP Hub: 23.9%, AbuseIPDB: 17.6%, IPVOID: 18.5%). Overall, though the results are similar, IP Hub appears to be comparable in terms of finding true positives, while being more accurate in locating true negatives, which is useful for practitioners in avoiding dealing with complaints from their workforce.

Figure A1: Data quality on retrospective study 1 using different IP traceback services



A3 R and Stata packages for interaction with IP Hub

This R package is already available on CRAN (<https://cran.r-project.org/web/packages/rIP/index.html>) and, in a more updated form, on GitHub (<https://github.com/MAHDLab/rIP>). For easy reference, we have copied the main part of the package into this appendix. This is the main function, not including the other parts that are required for R packages.

Figure A2: Snapshot of R package code for IP traceback

```
#' Passes an array of IP addresses to iphub.info and returns a dataframe with details of IP
#'
#' Makes a call to an IP address verification service (iphub.info) that returns the information on
the IP address, including the internet service provider (ISP) and whether it is likely a server
farm being used to disguise a respondent's location.
#'@usage getIPinfo(d, "i", "key")
#'@param d Data frame where IP addresses are stored
#'@param i Name of the vector in data frame, d, corresponding to IP addresses in quotation
marks
#'@param key User's X-key in quotation marks
#'@details Takes an array of IPs and the user's X-Key, and passes these to iphub.info.
Returns a dataframe with the IP address (used for merging), country code, country name,
asn, isp, block, and hostname.
#'@return ipDF A dataframe with the IP address, country code, country name, asn, isp, block,
and hostname.
#'@note Users must have an active iphub.info account with a valid X-key.
#'@examples
#id <- c(1,2,3,4) # fake respondent id's
#ips <- c(123.232, 213.435, 234.764, 543.765) # fake ips
#data <- data.frame(id,ips)
#getIPinfo(data, "ips", "MzI3NDpJcVJKSTdldXpQSUJLQVhZY1RvRxaXFsFW3jS3xcQ")
#'@export
getIPinfo <- function(d, i, key){
  if (!requireNamespace("httr", quietly = TRUE)) {
    stop("Package \"httr\" needed for this function to work. Please install it.",
      call. = FALSE)
  }
  if (!requireNamespace("utils", quietly = TRUE)) {
    stop("Package \"utils\" needed for this function to work. Please install it.",
      call. = FALSE)
  }
  if (!requireNamespace("iptools", quietly = TRUE)) {
    stop("Package \"iptools\" needed for this function to work. Please install it.",
      call. = FALSE)
  }
  #message("** Consider storing the ipDF as an object to write as an external df, e.g.,
```

```

write.csv(ipDF, 'ipDF.csv')
ips <- unique(d[,i])
options(stringsAsFactors = FALSE)
url <- "http://v2.api.iphub.info/ip/"
pb <- utils::txtProgressBar(min = 0, max = length(ips), style = 3)
ipDF <- c()
for (i in 1:length(ips)) {
  if(is.na(iptools::ip_classify(ips[i])) | iptools::ip_classify(ips[i]) == "invalid") {
    warning(paste0("Warning: An invalid or missing IP address was detected on line ", i, ".
Please check this.))
    next
  }
  ipInfo <- httr::GET(paste0(url, ips[i]), httr::add_headers(`X-Key` = key))
  infoVector <- unlist(httr::content(ipInfo))
  ipDF <- rbind(ipDF, infoVector)
  utils::setTxtProgressBar(pb, i)
}
close(pb)
rownames(ipDF) <- NULL
ipDF <- data.frame(ipDF)

return(ipDF)
}

```

Figure A3 shows a call to the R package. It takes as its input a data frame, the name of the variable containing the IP addresses, and the user's IP Hub key. The output is also a data frame, which includes the IP address. A user can use the data frame independently or merge it with their data by the IP address.

Figure A3: An example of the R package run in R terminal

```

> library(rIP)
> id <- c(1,2,3,4) # fake respondent id's
> ips <- c("196.19.158.224", "199.241.146.238", "190.124.31.51", "181.225.47.154") # fraudulent ips
> data <- data.frame(id, ips)
> getIPinfo(data, "ips", "MzI2MTpkovpld3pZTVg1VmdTV3ZPenpzMmhodkJmdEpIMkRMZQ==")
=====| 100%

```

| | ip | countryCode | countryName | asn |
|---|-----------------|-------------|-----------------------------------|-------|
| 1 | 196.19.158.224 | US | United States | 19969 |
| 2 | 199.241.146.238 | US | United States | 25780 |
| 3 | 190.124.31.51 | VE | Venezuela, Bolivarian Republic of | 61461 |
| 4 | 181.225.47.154 | VE | Venezuela, Bolivarian Republic of | 8053 |

```


```

| | isp | block | hostname |
|---|---------------------|-------|-----------------|
| 1 | JOESDATACENTER | 1 | 196.19.158.224 |
| 2 | HUGESERVER-NETWORKS | 1 | 199.241.146.238 |
| 3 | Airtek | 0 | 190.124.31.51 |
| 4 | IFX | 0 | 181.225.47.154 |

The Stata version of this package is available for download through SSC (<https://econpapers.repec.org/software/bocbocode/s458578.htm>) and can be installed by entering `ssc install checkipaddresses` into the terminal. Running the program consists of a single command line. Figure A4 shows the simple call to the package and re-merging with the user's data.

Figure A4: An Example of the Stata Package Run in the Stata Terminal

```
. [REDACTED] IP using ipresults.dta, xkey([REDACTED])
Checking 17 IP addresses:.....
iphub info saved; merge results into dataset with this command:
merge m:1 IP using ipresults.dta

. merge m:1 IP using ipresults.dta

Result                # of obs.
-----
not matched                0
matched                    18  (_merge==3)
```

Finally, some potential users of this package may be unfamiliar with R or Stata and, therefore, unable to use the package for their purposes. We created a Shiny application that can be used by anyone, regardless of software (e.g. SPSS, Stata, SAS, and Excel). The user can save their data as a .csv file using the “Save As” or “Export” feature in their software. The application takes that dataset and the user's key, and prompts them to enter the variable containing the IP addresses. When they press “Get IP Information”, the R package is called and a data frame with the IP information is displayed in the main panel. Finally, the user can click the button to download the data, as a .csv file, to their computer and can merge it with their dataset in any software. The online application is available for use at <https://rkennedy.shinyapps.io/IPlookup/>.

Figure A5: Online application for IP lookups

IP Lookup with IP Hub

IP Hub API Key (see instructions below):

Choose CSV file

Browse...
testData.csv

Upload complete

Column with IP Addresses:

IPAddress
▼

Get IP information
Download IP information as .csv

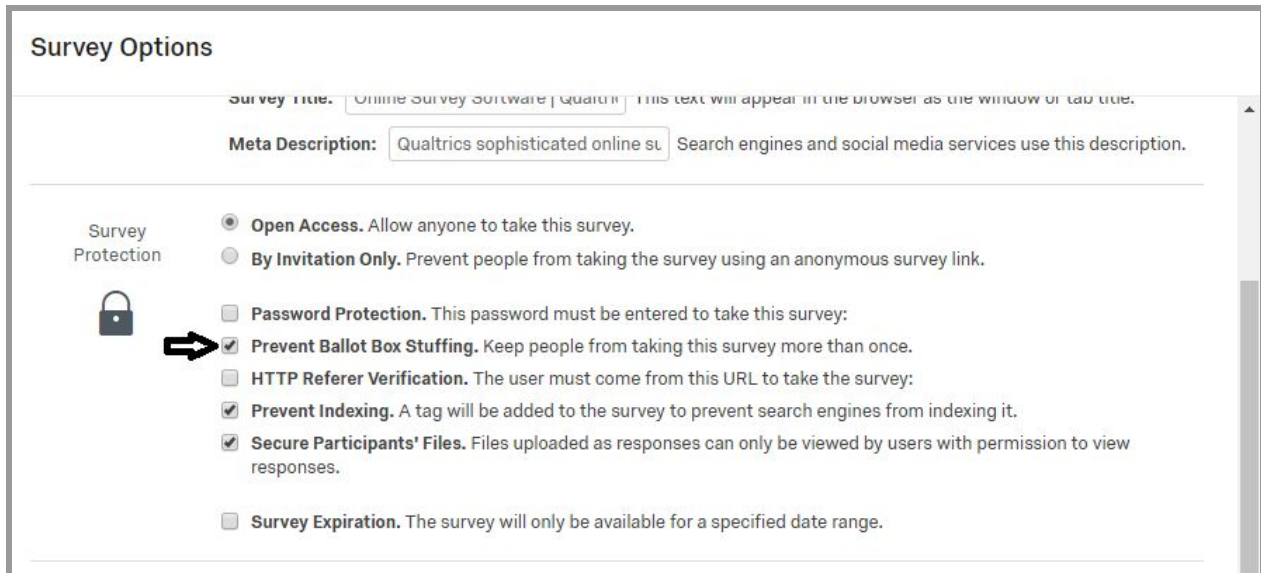
To get an IP Hub API key go to <https://iphub.info/api>. A free account is available that allows 1,000 lookups per day. Once you have entered your key, upload your data. The data must be in .csv format. You can save your data in this format from just about any statistical software (including Excel, SPSS, Stata) by either using "Save As" or "Export". Similarly, the downloaded .csv file can be uploaded into your preferred software using "Open" or "Import".

| ip | countryCode | countryName | asn | isp | block | hostname |
|-----------------|-------------|---------------|-------|---------------------------|-------|-----------------|
| 24.98.143.237 | US | United States | 7922 | COMCAST-7922 | 0 | 24.98.143.237 |
| 98.28.120.247 | US | United States | 10796 | SCRR-10796 | 0 | 98.28.120.247 |
| 208.58.123.127 | US | United States | 6079 | RCN-AS | 0 | 208.58.123.127 |
| 24.116.56.240 | US | United States | 11492 | CABLEONE | 0 | 24.116.56.240 |
| 98.245.66.27 | US | United States | 7922 | COMCAST-7922 | 0 | 98.245.66.27 |
| 76.23.61.247 | US | United States | 7922 | COMCAST-7922 | 0 | 76.23.61.247 |
| 70.176.4.129 | US | United States | 22773 | ASN-CXA-ALL-CCI-22773-RDC | 0 | 70.176.4.129 |
| 68.50.26.34 | US | United States | 7922 | COMCAST-7922 | 0 | 68.50.26.34 |
| 174.218.19.51 | US | United States | 22394 | CELLCO | 0 | 174.218.19.51 |
| 172.221.187.108 | US | United States | 20115 | CHARTER-NET-HKY-NC | 0 | 172.221.187.108 |

A4 Qualtrics screening protocol for blocking VPS and international respondents

This protocol takes you through the steps of setting up a filter on Qualtrics that will block most people in a non-US location¹ or using a Virtual Private Server (VPS) to cover their location. It will take you through all the steps, with illustrations as needed. It is also available on SSRN (<https://papers.ssrn.com/abstract=3327274>).

Before you begin, if you do not want people to be able to answer your survey more than once, you should always enable the Prevent Ballot Box Stuffing option in Survey Options.



The screenshot shows the 'Survey Options' page in Qualtrics. At the top, there are fields for 'Survey Title' and 'Meta Description'. Below this is the 'Survey Protection' section, which includes a lock icon and a list of options:

- Open Access.** Allow anyone to take this survey.
- By Invitation Only.** Prevent people from taking the survey using an anonymous survey link.
- Password Protection.** This password must be entered to take this survey:
- Prevent Ballot Box Stuffing.** Keep people from taking this survey more than once.
- HTTP Referrer Verification.** The user must come from this URL to take the survey:
- Prevent Indexing.** A tag will be added to the survey to prevent search engines from indexing it.
- Secure Participants' Files.** Files uploaded as responses can only be viewed by users with permission to view responses.
- Survey Expiration.** The survey will only be available for a specified date range.

This option places a cookie on the user's browser to prevent them from answering the survey more than once from the same browser. It does not completely prevent duplicate responses, since users can take steps to erase or avoid detection through cookies. But it is a useful supplement to MTurk's built-in checks to avoid duplication.

¹ This can also be used to select only participants from other countries by simply changing the IP_countryCode parameter.

Once this is done, you can follow the following steps to detect international and VPS respondents. The main concept is to lookup the IP address using a security service and use that information to make decisions on how to handle the potential respondent.

1. Create an account on IP Hub (<https://iphub.info/pricing>). A free plan that allows for 1,000 requests per day should suffice for most research purposes, although larger plans are available. We recommend IP Hub based on our own experiences, and because it provides a relatively liberal free service that functions quickly. You will be given an API key that consists of about 50 random letters and numbers, looking something like this:²
MxI5ODpZT2kmVnlsR5iMcjBrRWpjxVZOKXIRKU1sNmdZb30EMA==
2. Next, you will want to be sure to add a warning to the beginning of the survey to tell people who are in the U.S. to turn off their VPNs or any ad blocking software they are using. This should be placed in its own block and should come before any other parts of the survey. This will prevent you from receiving complaints from some Turkers. From our piloting, it also appears that this is an effective way to initially screen out people who you do not want to take the survey (we noticed a significant drop in the number of international IPs testing our system once we added the warning).

Warning!

This survey uses a protocol to check that you are responding from inside the U.S. and not using a Virtual Private Server (VPS), Virtual Private Network (VPN), or proxy to hide your country. In order to take this survey, please turn off your VPS/VPN/proxy if you are using one and also any ad blocking applications. Failure to do this might prevent you from completing the HIT.

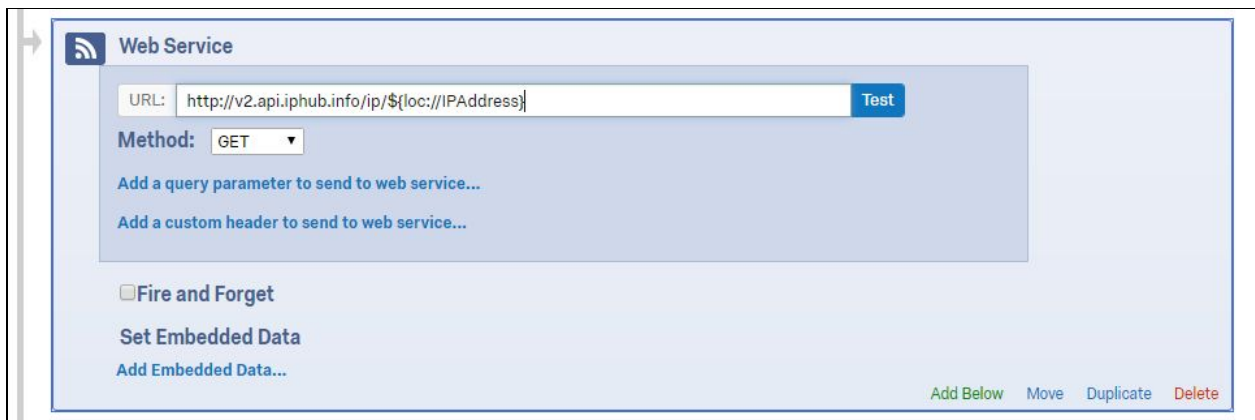
For more information on why we are requesting this, see this post from TurkPrime (<https://goo.gl/WD6QD4>)

² This is not a real key, please do not try to use.

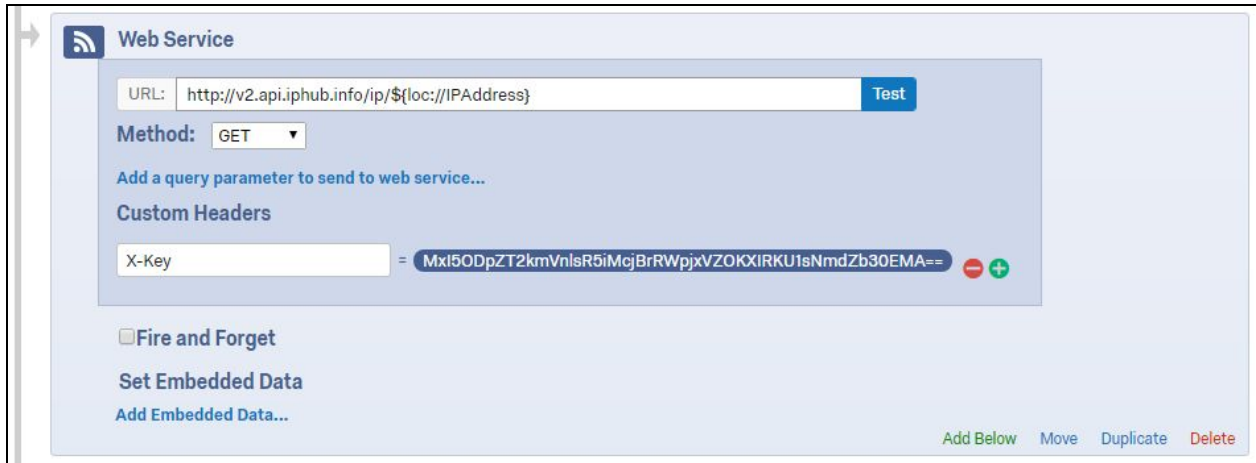
- Next, go to the Survey Flow of your Qualtrics survey. After the block that contains the warning, you should add a Web Service.



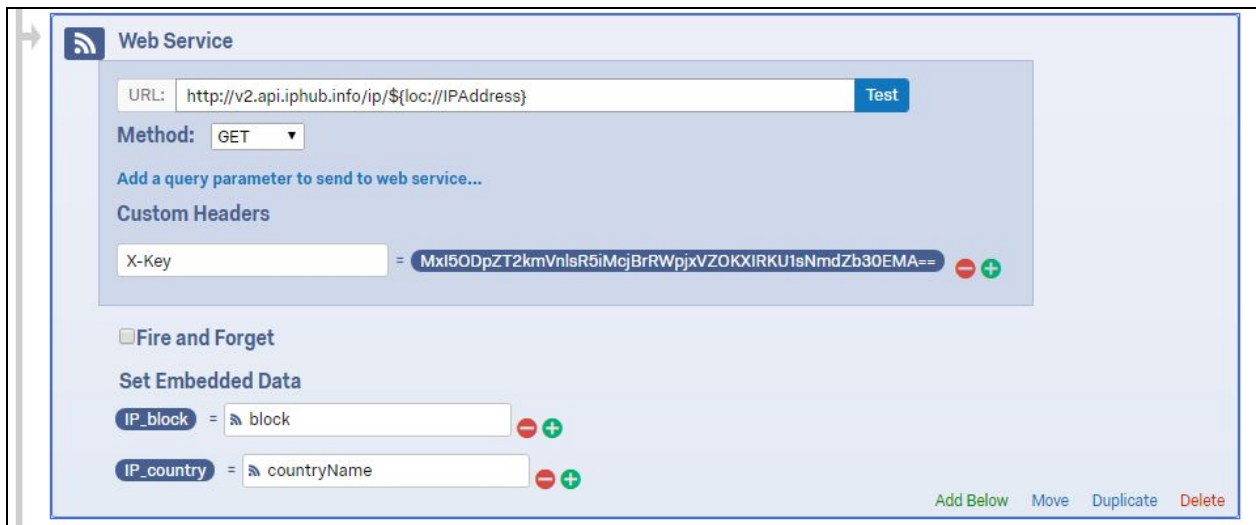
- This is where we will make the call to IP Hub. In the "URL:" line, place in the following address: `http://v2.api.iphub.info/ip/${loc://IPAddress}`. The first part of this address calls the IP Hub API, the last part takes the IP address captured by Qualtrics and adds it to the API call. Make sure the "Method:" is set to "GET".



- Click on “Add a custom header to send to web service...” On the left-hand side, for “Header to Web Service...” type in X-Key. On the right-hand side, where it says “Set a Value Now”, type in your API key.



- Next, click on “Add Embedded Data...” and add entries for IP_block and IP_country that correspond with the returned IP Hub fields block and countryName.



7. Click Save Flow. If you would like to screen out international or VPS actors, you can do this using the next few steps. Begin by setting up warnings explaining why they are not allowed to take the survey. These can be added as descriptive text questions in their own Block (or text entry questions if using the appeals procedure described at the end). This is both courteous and will prevent you from getting nasty emails. Below are the ones we used.

VPS Warning:

Qualtrics View

▼ VPS Warning Block Options ▼

VPS

Our system has detected that you are using a Virtual Private Server (VPS) or proxy to mask your country location. As has been widely reported, this has caused a number of problems with MTurk data (<https://goo.gl/WD6QD4>).

⚙️ Because of this, we cannot let you participate in this study. If you are located in the U.S., please turn off your VPS the next time you participate in a survey-based HIT, as we requested in the warning message at the beginning. If you are outside of the U.S., we apologize, but this study is directed only towards U.S. Participants.

Thank you for your interest in our study.

Survey View

Our system has detected that you are using a Virtual Private Server (VPS) or proxy to mask your country location. As has been widely reported, this has caused a number of problems with MTurk data (<https://goo.gl/WD6QD4>).

Because of this, we cannot let you participate in this study. If you are located in the U.S., please turn off your VPS the next time you participate in a survey-based HIT, as we requested in the warning message at the beginning. If you are outside of the U.S., we apologize, but this study is directed only towards U.S. Participants.


Thank you for your interest in our study.

Out of US Warning:

Qualtrics View

Out of US Warning Block Options

OutofUS Our system has detected that you are attempting to take this survey from a location outside of the U.S. Unfortunately, this study is directed only towards participants in the U.S. (this also excludes U.S. citizens living abroad) and we cannot accept responses from those in other countries (as per our IRB protocol).

 Thank you for your interest in our study.

Import Questions From... + Create a New Question

Survey View

Our system has detected that you are attempting to take this survey from a location outside of the U.S. Unfortunately, this study is directed only towards participants in the U.S. and we cannot accept responses from those in other countries (as per our IRB protocol).

Thank you for your interest in our study.

Still Missing Warning

(this message is added defensively. We find a small number of cases (about 1.6% in our pilot) the API lookup does not succeed and responses need to be checked after the survey is complete):

Qualtrics View

Still Missing Warning Block Options

WID For some reason we were still unable to verify your country location. We ask you to please assist us in getting this protocol correct. Please enter your MTurk worker ID below and contact the requester for this HIT to report the problem.

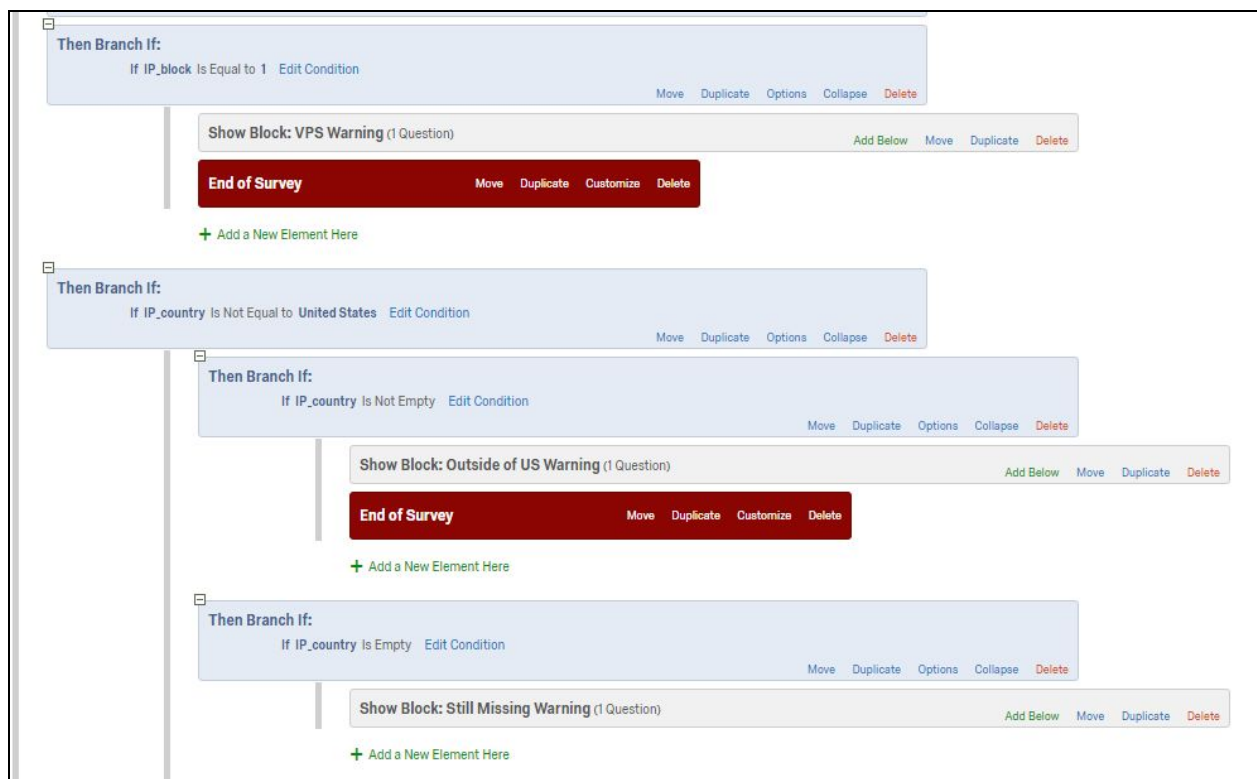
Once you click Next, you will be taken to the survey (and certifying that you are taking this survey from the U.S. and not using a VPS). We will be checking locations manually for those who reach this point and you will be contacted if this check identifies you as violating these requirements.

Survey View

For some reason we were still unable to verify your country location. We ask you to please assist us in getting this protocol correct. Please enter your MTurk worker ID below and contact the requester for this HIT to report the problem.

Once you click Next, you will be taken to the survey (and certifying that you are taking this survey from the U.S. and not using a VPS). We will be checking locations manually for those who reach this point and you will be contacted if this check identifies you as violating these requirements.

8. Now go back to the Survey Flow. After the web service call we added earlier, add two Branches that respond to Embedded Data. For the first one, set it to activate “If IP_block is Equal to 1”. Move your VPS warning text underneath this branch and then add an End of Survey option below it. For the second Branch, set it to activate “If IP_country is Not Equal to United States”, then create two sub-Branches for “If IP_country is Not Empty” and “If IP_country is Empty”. Drag your out of US warning underneath under the first sub-Branch and add an End of Survey option below it. Under your second sub-Branch where IP_country is empty (this means that there was an error in the IP trace) drag your location missing warning. Now if anyone tries to access your survey from outside the US or from a server farm, they will be shown a warning and taken to the end of your survey. This part of your survey flow will look like the illustration below.



9. Save the new Survey Flow. Now nobody (or at least very few people) outside of the US or using a detected server farm should be able to take your survey.

[One final note. With any IP lookup service, there is always a risk of false positives -- people who get screened out when they should not have been. So far, we have only encountered two cases of this. But, because it can happen, and because those who report false positives are helping us evaluate the utility of the screening procedure, we also included an appeal process for Turkers. We added to the information screens shown above the line, "If you have received this message in error, please contact the requester and enter your MTurk worker ID in the box below" and added a text box for entry (see illustration below). In our case, the information provided to us about false positives from these respondents was worth paying them what they would have received for the survey, but this is left to the discretion of the researcher. With Amazon continuing to update their service and remove fraudulent participants, it is possible that one day the number of false positives will supersede the utility of this protocol, but we have, so far, continued to find it useful.]

VPS Exclusion Message with Appeal Process

Our system has detected that you are using a Virtual Private Server (VPS) or proxy to mask your country location. As has been widely reported, this has caused a number of problems with MTurk data (<https://goo.gl/WD6QD4>).

Because of this, we cannot let you participate in this study. If you are located in the U.S., please turn off your VPS the next time you participate in a survey-based HIT, as we requested in the warning message at the beginning. If you are outside of the U.S., we apologize, but this study is directed only towards U.S. Participants.

Thank you for your interest in our study.

If you have received this message in error, please report it to the requester for this study and enter your MTurk Worker ID below.

A5 A note on human subjects compliance

Both the use of the “rIP” package and the screening protocol require the monitoring of IP address information. Some research institutions, including all institutions in Europe and governed by the General Data Protection Regulation (GDPR), will consider this personal identifying information (PID). This may mean that the use of these protocols may require a full review by the researcher’s Institutional Review Board (IRB). In this section, we provide some sample wording for presenting these issues to your IRB if this is required. We present a sample question followed by a sample response.

Question: Does your study collect PID? If so, why?

Answer: Yes, our study does record the respondent’s Internet Protocol (IP) address as defined by [this institution/the GDPR], a numerical label assigned to each device connected to a computer network that used an Internet Protocol for communication. This is defined as PID by this institution. It is collected in order to determine whether respondents are misrepresenting their location in a manner that would disqualify them from participating in the study under our stated selection criteria, which stipulates that respondents must be located in [COUNTRY OF DESIRED RESPONDENTS]. Several studies have found that online survey participants will sometimes misrepresent their location and use Virtual Private Servers (VPS) to mask their location ([Dennis et al. 2018](#); [Ahler et al. 2018](#); [TurkPrime](#) ; [Kennedy et al. 2018](#)). Including these respondents could both harm the validity of results from the study and would mean inclusion of respondents outside of the IRB approved study population. The only reliable method for mitigating this problem is to collect IP addresses and check them against a database of known VPS service providers and identify the country of origin for the IP address.

Question: How does your study reduce potential harm from collection of PID?

Answer [if using “rIP”]: We utilize an R package called “rIP” to evaluate the IP addresses ([Waggoner et al. 2019](#)). This package interacts with the APIs of several databases and returns only information about the country of origin of the IP address, whether it is likely outside of the area of interest for the researcher, and the internet provider. By returning only this information, it does not provide any information that could be used for further identification of the respondent by the research team. This information is also not sufficient, in combination with other data collected, to make a personal identification. Any dissemination of this data will only include a dichotomous variable indicating whether the observation was marked as likely being from a VPS or a country outside the purview of the study. It will not include IP address or any additional information derived from the IP address.

Answer [if using screening protocol]: To avoid collecting surveys from individuals outside our stated target population, we utilize a screening protocol based on IP address ([Winter et al. 2019](#)). We warn all survey respondents that we are only collecting data from the specified country in our human subjects protocol, and tell them to turn off their VPS before proceeding to the survey. We utilize an IP lookup service called IP Hub to check whether the IP address is from a VPS or from outside the target country. The only information recorded is the country of origin and whether the server is in the IP Hub database as a VPS. Those who are detected to

be outside the recruiting protocol are taken directly to the end of the survey and provided an explanation for why they are not eligible to participate in the study. [We provide them with contact information they can use if they believe an error has occurred and would like to take the survey.] Any dissemination of this data will remove both the IP address and all information collected from IP Hub.

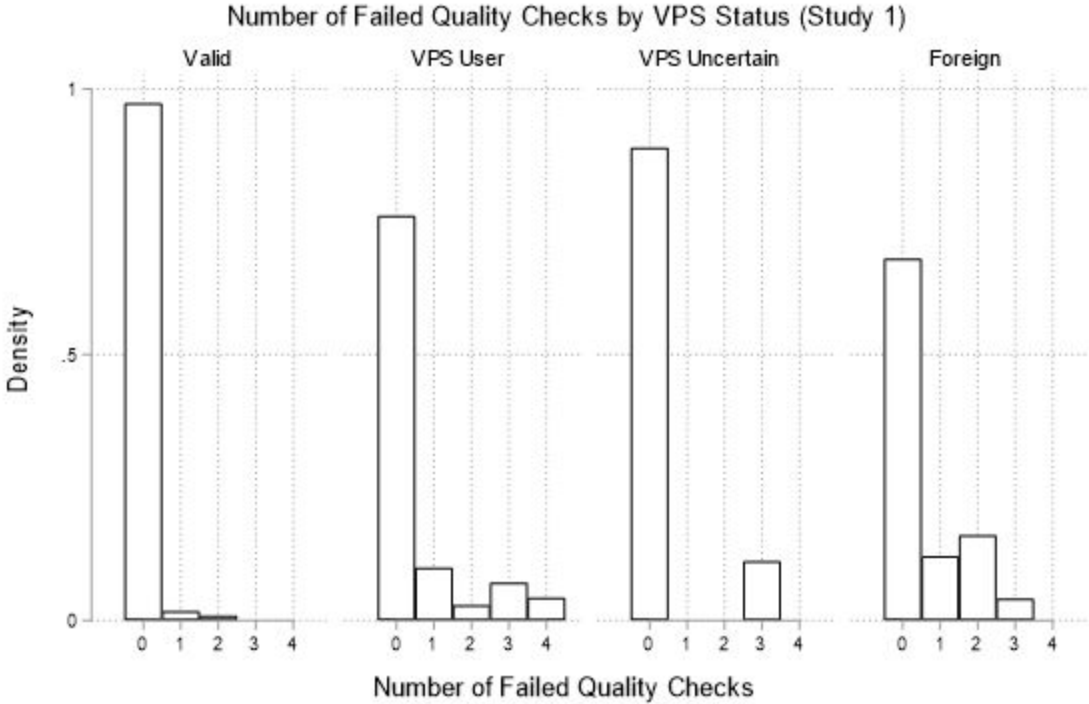
A6 An analysis of individual quality checks

The table below displays the failure rates and correlations between the five quality checks included in Study 1. Failure rates are rather low among the full sample. The quality checks tend to be positively correlated with each other, but these associations tend to be weak. However, the three quality checks that relied on open-ended text (location, purpose, comments) tend to be strongly correlated with each other.

Correlations between quality checks in Study 1

| | % Failing | City | Age | Location | Purpose | Comments |
|----------|-----------|-------|------|----------|---------|----------|
| City | 1% | - | | | | |
| Age | 4% | 0.10 | - | | | |
| Location | 2% | 0.15 | 0.20 | - | | |
| Purpose | 4% | 0.11 | 0.40 | 0.76 | - | |
| Comments | 2% | -0.01 | 0.16 | 0.66 | 0.70 | - |

The histogram below displays the number of quality checks that respondents failed in Study 1 by user IP status. For all IP categories, the modal number of failed checks is zero. Among those who failed at least one check, the modal number of failed checks is one for all IP categories, with the exception of foreign IPs, who were more likely to fail two.

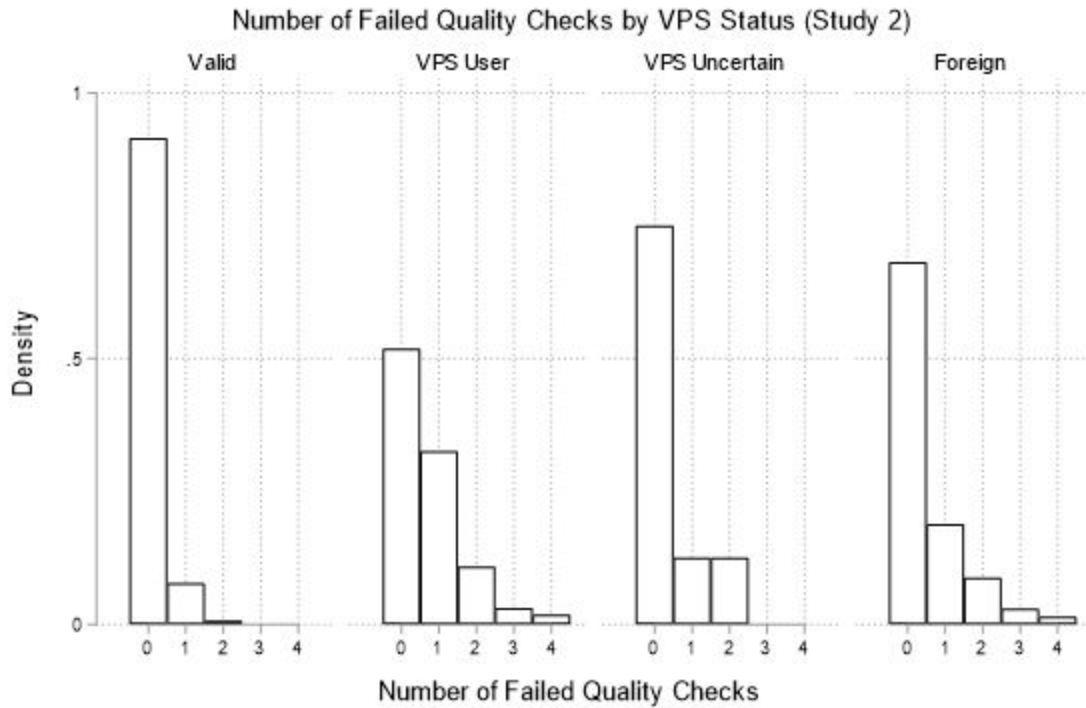


The table below displays the failure rates and correlations between the five quality checks included in Study 2. Failure rates are rather low among the full sample. The quality checks tend to be positively correlated with each other, but these associations tend to be weak. In contrast to Study 1, however, the three checks based on open-ended comments are not clearly more strongly related to each other.

Correlations between quality checks in Study 2

| | % Failing | City | Age | Location | Purpose | Comments |
|----------|-----------|------|------|----------|---------|----------|
| City | 1% | - | | | | |
| Age | 4% | 0.23 | - | | | |
| Location | 5% | 0.07 | 0.02 | - | | |
| Purpose | 35% | 0.19 | 0.18 | 0.08 | - | |
| Comments | 2% | 0.16 | 0.18 | 0.13 | 0.33 | - |

The histogram below displays the number of quality checks that respondents failed in Study 2 by user IP status. For all IP categories, the modal number of failed checks is zero. Among those who failed at least one check, the modal number of failed checks is one for all IP categories.



A7: Analysis of ISP and timing of survey responses

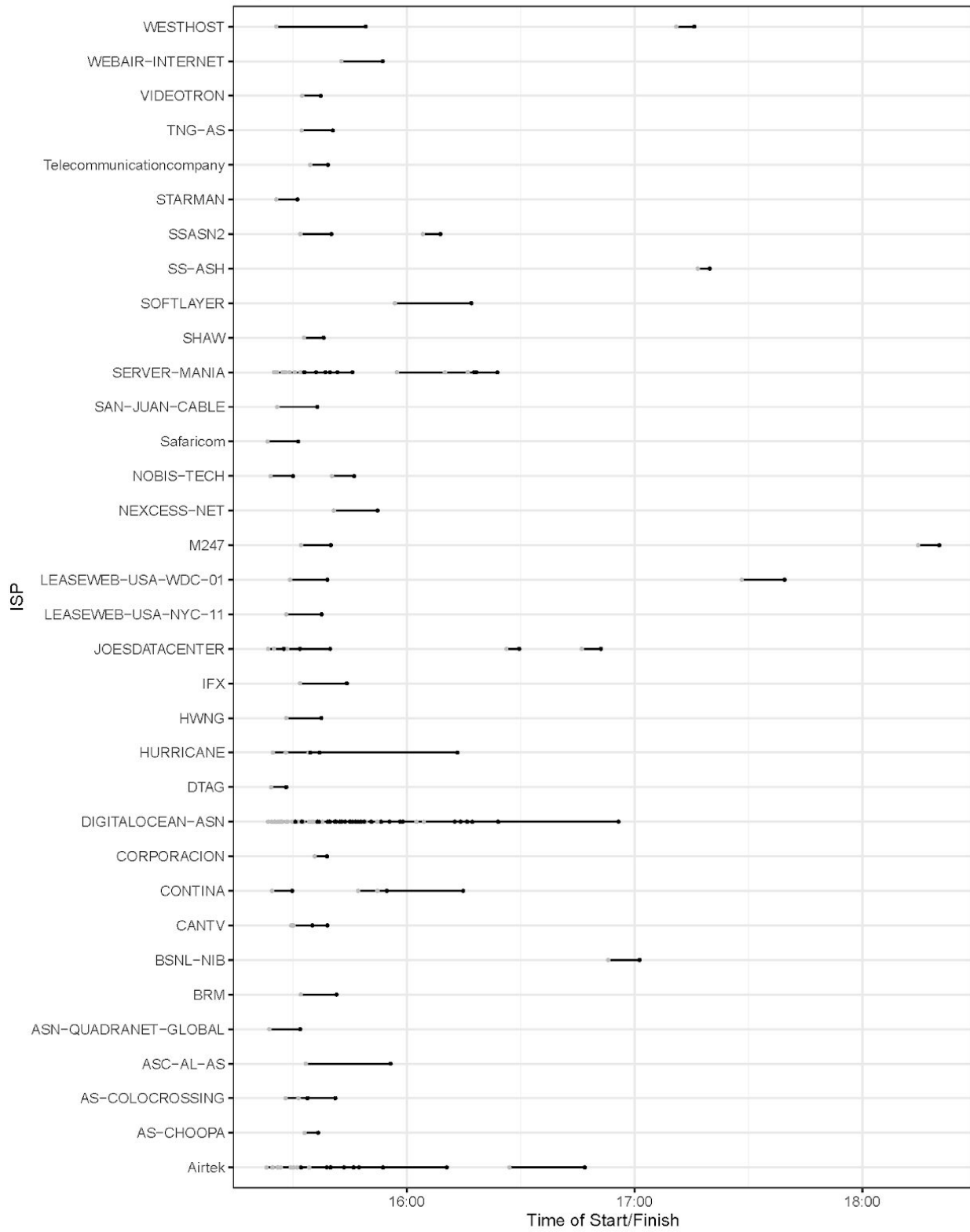
An important question impacting how to deal with the problem identified in this paper is whether this is just a handful of people who are taking the surveys using a number of different MTurk IDs, or whether there are a number of people who participated in this. While it is difficult to say which is correct with certainty, we conduct a couple of analyses below to look into this question. The figures below show dumbbell plots for the start and end time of each survey response for Experiment 1 and Experiment 2, divided by Internet Service Provider (ISP). The beginning time is in grey, and the end time is in black.

The charts can be a little difficult to read because of overlaps, especially with popular VPS services like DigitalOcean. We should also note that they are different between the two experiments, in part because Experiment 1 was done as one wave, while Experiment 2 was fielded in two waves.

Nonetheless, there are several patterns that are apparent from the chart, and we believe these patterns suggest that there were quite a few people involved in the behavior identified in the main paper. First, there are a relatively large number of different ISPs listed. It seems unlikely that a handful of people would switch their ISPs so regularly and use both VPS and non-VPS ISPs in their responses. Second, there is too much overlap between submission times. The timing of surveys taken from the same ISP overlap quite a bit and show a quite a bit of variety in terms of time taken to complete the surveys. These overlaps are both within and between ISPs. Third, we see a number of ISPs in both experiments with only one response. This does not fit well with this just being a handful of people, since, to reach the needed scale, switching ISPs for a single response is not very efficient.

As noted, this is not dispositive that at least some of the problem may be due to a smaller group of fraudulent respondents, but it does strongly indicate that the problem is not limited to a handful of users.

Experiment 1



Experiment 2

