Retirement Information and 'BS' Detection in Online Communities DRAFT

Raymond Duch
raymond.duch@nuffield.ox.ac.uk
Nuffield College
University of Oxford

Sönke Ehret
sonke.ehret@nuffield.ox.ac.uk
Nuffield College
University of Oxford

January 10, 2018

*Paper prepared for presentation at the 2018 Asia Political Methodology Society Meeting, Seoul, South Korea January 11-12, 2018.

We adopt two strategies in order to isolate the information content or frames that signal "Bullshit" to the average individual. One is simply observational: We observe a very large corpus of financial information postings to the Reddit social media platform (specifically those threads concerned with retirement investments); these postings are scraped and their content coded; we then observe (and content analyze) the resulting commentary from the Reddit community along with their positive/negative vote for the posting.

A second element of the design exploits random assignment to help isolate the causal mechanism – the specific feature of the information content that triggers a positive response. We randomly post information treatments again to the same Reddit investment treads; we then content analyze the resulting commentaries and the votes.

There were four important results from phase one of the analysis of the observed corpus of postings and comments from the retirement investment threads on Reddit: 1) sentiment analysis was surprisingly accurate allowing us to easily categorize the negative and positive reactions to the posting in line with up and down votes of posts; 2) topics are unambiguously identified by both human coders and automated topic modeling – for example a topic based on promises of free income/credit 3) our automated bullshit modeling using machine learning techniques (random forests) was very accurate; 4) as was our "fake information" modeling using similar approaches.

The observational phase provided four key features of bullshit and fake news that facilitate detection by the Reddit community: 1) key words indicating 'BS', 2) language complexity which we expect to positively correlate with BS, 3) purposeful dissonance/ conflict detection on behalf of subjects, and 4) domain expertise. (Results on 3) and 4) are not reported here.) On the other hand, there are four aspects of bullshit and fake news that make them almost impossible to detect: 1) BS is not obviously wrong, but 2) creates an impression of profoundness and 3) masks its lack of substance through incomprehensible statements, wording and sentence structure. In particular, it appears that redditors are prone to accepting and even promoting 'BS' posts, even when said redditors are engaged and experienced members of subreddits on financial decision making.

The second part of our study incorporates random assignment to treatments –

the experiments were designed to assess in a causal fashion the impact of different content and framing on the ability of the Reddit community to detect fake news and bullshit. The treatments build on the findings from the Phase one observational study. Drawing on the existing data, dictionaries of BS are created through human coders using two methods, one to rate posts independently on two dimensions of profoundness and comprehensibility, and two by using pre-selected 'BS' terms compiled by the researchers and crowdcoders.

The design is relatively complicated – we implement a 4 X 10: control-null, control-placebo; bullshit/fake; true. We then vary 10 features of the content/framing that we expected to signal the authenticity of the information. In total we implement 40 different treatments in order to help identify the characteristics of online information that facilitates and undermines the ability of consumers to detect fake news and bullshit, that is, to make judgments on BS texts where both profoundness and intelligibility are represented equally. In real detection scenarios, participants have to both decide on both dimensions at the same time. Six of these treatment effects appear to be particularly powerful in enabling participants to reflect on the incomprehensibility of BS statements.

1 Motivation

The rise of social media has created renewed concerns about "fake news" and their effect on the ability of average citizen media consumers to discern, evaluate and if necessary, reject information of dubious or objectively false origin. These concerns of the impact of false information cover a broad range of decision making – financial; consumption good; politics; health; and pop culture.

In this essay we focus on financial decision making by the general public. There are of course a broad range of financial decisions that individuals and households make on a regular basis. And there is a burgeoning literature, much of it from behavioral economics, that documents the extent to which consumers make sub-optimal financial choices.

Over the past couple of decades research in behavioral economics has clearly established that the heuristics and biases that shape individual decision making can lead to sub-optimal financial choices (Tversky and Kahneman 1974, Thaler

and Sunstein 2009). Households are prone to make a variety of investment mistakes such as lack of diversification, risky share inertia, and a disposition to hold losing and sell winning stocks (Calvet, Campbell and Sodini 2009). And there is evidence that households make systematic errors over time that significantly affect wealth accumulation (Ameriks, Caplin and Leahy 2003). There is evidence of a U-shaped relationship between age and financial mistakes –(Hastings, Madrian and Skimmyhorn 2013) suggest, for example, that with age individuals accumulate increased financial decision-making abilities and then after the age of fifty these financial reasoning abilities decline. (Mani et al. 2013) provide persuasive evidence that poverty represents an important inhibitor of cognitive functioning and in particular can lead to poor financial decision making.

We have also made advances in our understanding of what causes individuals to make better or worse financial choices. Social networks can reduce the costs of becoming informed about optimal retirement investment vehicles, for example (Duflo and Saez 2003). Efforts to design the provision of information to consumers and the regulation of financial service providers build on these causal insights into financial decision making by consumers.

The rise of social media has complicated this task. First, financial decisions, such as those concerning retirement, are increasingly being informed by, and are being taken on, the Internet. Second, social media is increasingly the source of 'fake news' and 'bullshit'. And as regards to retirement decisions, financial advice, fake advice, or the closely related concept of "bullshit" advice can have grave consequences on the average citizen investors.

The Internet and social media give anybody with the commercial intention or fraudulent aptitude to give such misleading advice on investment choices a platform. The multitude of potential sources makes the detection of nonsensical but "pseudo-profound" financial information or fake advice very difficult. Potential social media filtering effects, the concentration of within-group dynamics and echo chambers worsen the problem. For individual citizen investor the problem is therefore compounded, since average financial literacy to make informed retirement choices is low. Experimental evidence clearly suggests that individuals have a difficult time making even the most basic financial decisions. Given their low epistemic capacity and also capability, citizen investors will rely on cues and other

information shortcut heuristics, that is, on forms of "low level", fast and intuitionist thinking. As prior studies have pointed out, fake news and profound nonsense may have detrimental effects due to these low level cognitive responses.

However whether "heuristics" lead to inferior outcomes in financial decision making or not is a function of the micro-social context in which they are employed. Citizen investors face significant stakes. While it is possible that financial decision making is entirely led by consumers buying into the latest fads and risky investments, it is also important to understand whether the heuristics employed in financial decision making are equally likely to be of a critical nature.

There are two underlying issues:

1) What is the heuristic of high stake environments? Does it provide tools for "bullshit" detection as well and bias towards conservative retirement choices? We cannot be sure whether participants are more likely to commit type I detection error and falsely believe a product to be effective, or commit type II error and reject a working financial product due to conservative heuristics.

2) Will time and process lead participants to 'conflict' detection at all, i.e. finding features of an argument that make them second guess the argument, or not. Here we care about the psychological process. The classical two step intuitive-deliberative model in which conflict detecting deliberation supersedes intuition in linear fashion over time appears to be context specific, that is, not universal. Our research implements treatments that are designed to encourage conflict-detection; and, alternatively, to make detection more difficult. We expect that this will provide tighter control over the psychological mechanism of BS detection in large scale deliberative communities.

Research approaches. With the rise of fake news and concern about its impact on citizen decision making there has been an increase in research designed to better understand the phenomenon. Much of the existing research simply characterizes the prevalence of fake information particularly on social media (Allcott and Gentzkow 2017). And the evidence clearly confirms the growing prevalence of fake information.

But does it matter? Is BS and Fake News detected by average consumers? Or, as we speculated above, are average consumers ill-equipped for detecting this false information? And is some BS and Fake News more easily to detect than others? The causal claim here is relatively straight forward: The outcome of interest there is a "correct" evaluation of a message. One conjecture is that consumers are fully informed and hence there is a strong causal relationship between the authenticity of a message and the subject's assessment of its authenticity. Concern about fake news and bull shit is based on an hypothesized weak causal relationship between message authenticity and the subject's assessment of its authenticity.

There is limited rigorous causal evidence linking message authenticity to subject acceptance (or rejection if its not authentic). As we hinted to above, the process by which individuals consume and digest information on social media is probably not a simple instantaneous response to a posted message. It could be but our intuition is that there is more social interaction and deliberation that occurs particularly with information that has significant financial implications.

There are decision theoretic experiments or other singular non-interactive studies exploring the ability of individuals to detect bullshit or fake news. Included here are survey experiments with simple framing (Berinsky 2017) and MTurk survey experiments (Pennycook and Rand 2017, Pennycook, Cannon and Rand 2017).

These studies limit our ability to understand financial "bullshit" detection and its antecedents. Much of real world discussion for example with the partner, the family or friends can quickly reveal whether a scheme is nonsensical or not. That is, non-interactive experiments would not inform us about natural social correctives nor would they provide us the means of discerning the dynamics, learning and speed of information evaluation in the long run since they are essentially based on idiosyncratic samples without variation of external context.

To overcome the epistemic limitations of survey experiments, we need to employ both larger datasets on discussions on potentially profound but nonsensical financial information and also have experiments on the discursive constructs average citizens are using to dissect information. If the receptivity to fakery relies on fast thinking, to be corrected by slow deliberative processes (again, we can only assume), then online communities can provide remedies as well due to their sequential and discursive nature. Importantly then, how and when do heuristics

and deliberation interact, and in which way? Both in terms if variation of context as well as in terms of ecological validity the Internet can provide us with a viable laboratory for finding out the drivers of bullshit detection in a crucial domain of decision making.

Fake News and Bullshit

Fake news is not contested news per se, as when conflicting opinions are presented, or else, information that is rejecting established authority. Fake news is a provision of information that is willfully promoting false facts, i.e. it is information that intentionally reverses and falsifies and reverts our best guess on an uncertain state of the world without providing a verifiable, unambiguous and repeatable method to verify claims made.

In contrast to fake news, bullshit is profound sounding but nonsensical information that is orthogonal to the true state of the world, i.e. bullshit is not a masking of the truth nor a mistaken presentation of ???true??? information but is regardless of information a means to convey a sense of structure, story or information using socially acceptable and profound sounding language.

There is an emerging literature on fake news and bullshit. Yet this literature is mostly decision theoretic and uses small samples that are high on internal but low on both external and ecological validity. Furthermore, these findings have suggested that 1) intuitive heuristics are generally problematic and a root cause of financial wrong doing of average, undereducated individuals, and 2) that the Internet and social media are a means to trigger and deepen these inefficient heuristics. While our approach affirms the basic intuitions behind this reasoning we also seek to extend the scope of present research by highlighting the domain of financial decision making and the importance of discursive as opposed to non-discursive social media.

Detecting and Defining Bullshit.

Bullshit Dictionary Hence an initial phase of this project involved the development of a dictionary of financial Bullshit. We employed two strategies for the identification of social media content that could unambiguously be identified as bullshit. We began with the identification of a sample of postings to Reddit financial forums – these are essentially a subset of our total corpus of scrapped Reddit posts.

Our initial coding strategy is to have two different sets of MTurk coders provide the following scoring:

- One set of raters just evaluates whether a comment is 'profound' and 'impressive' on a five item scale.
- A second set of raters judges whether a post is intelligible and they understand it, again on a five item scale.

Postings are scored as BS if they are high on impressiveness but low on being understandable.

These coding exercises generated a comprehensive dictionary of financial terms along with bullshit and fake news scores that would be used for the dictionary (wordcount) coding of Reddit financial postings.

Evaluating fakery receptivity using observational online data

The Reddit Postings. We obtained the complete set of Reddit postings for the period 2007 to 2017. The first step in the analysis was identifying the subreddits that were finance-related. We matched subreddit names and brief descriptions against a standard dictionary of financial identifiers. Matched subreddits were included in our corpus. Our first step was simply to scrape all posts and comments from finance related subreddits from 2008 to 2015.

We have assembled all postings on Sub-reddit discussion forums that deal with issues related to personal finance. These data include all posts and comments for the period 2008 to 2015. We have a total of 212,000 posts and 3,120,000 comments.

posts	comments	authors	net-votes
211919	3127087	297877	Min. :-2404, Median : 1, Max. : 5619

The Reddit Metrics Each of these postings receives a Reddit-related metric. These included: 1) the average measure of approval in the form of net positive-negative votes; 2) whether the post has been cross-posted or not / shared or not on Reddit; 3) pending data availability one could also look at moderators and whether a post gets removed from moderators on a subreddit; 4) the change of 'karma' associated with the individual redditors posting; and 5) the velocity of which a post is shared or cross posted. All these variables are already contained in the dataset.

Crowd-sourced Coding of Reddit Postings. We completed the pilot crowd-sourced coding of 1000 posts, each post rated three times. Coders are shown the title, the initial post and a randomly selected comment.[1] We solicit a set of two ratings by independent raters. The crowd-sourced workers are asked to rate the profoundness of the post (whether it appears to be insightful and knowledgeable) and whether the post is intelligible.

Table 1: Crowdworker Ratings (%)

	Not profound	2	3	4	Very profound
Not comprehensible	3.00	1.25	**0.75**	**1.00**	**0.75**
2	4.00	5.50	**2.50**	**5.50**	**2.25**
3	1.75	2.00	**0.50**	**2.25**	**1.00**
4	5.00	7.25	7.00	8.75	2.75
Very comprehensible	7.25	9.00	3.75	9.25	6.00

Short binary coding (not profound and/or intelligible)

	Not profound sounding	profound sounding
Not comprehensible	21.25	**12.75**
comprehensible	39.25	26.75

A second step was to use these expert suggestions to build a dictionary of financial BS terms. Examples of the expert BS terms including the following: cross-

[1] we restrict the sample to those posts that receive comments (2.850.000

platform, bricks-and-clicks, transparent, cutting-edge, user-centric, visionary, convergence, deliverables, finance-markets, functionalities, initiatives, asset management customized, ubiquitous, collaborative, compelling, holistic, rich, adviser, advisers, analysts, adjustment, brokers, options, institutions, trusts, corporations, exchanges, trades, quotes, communities, convergence, deliverables, finance-markets, functionalities, initiatives, asset management, partnerships.

Table 2: BS-count

	count of posts with BS term
0	2439307
1	306630
2	72023
3	21940
4	8506
5	3846
6	1819
7	917
8	579
9	351
10	173
11	112
12	78
13	52
14	39
15	22
16	16
17	19
18	17
19	9
20	4
21	5
22	1
23	6
24	5
25	3
26	1
27	1
28	1
30	1
34	1
35	1
37	1
38	1
40	1
43	1
58	1
59	1

Flesch Complexity Score. Linguists have developed a range of readability measures. The Flesch-Kincaid[2] measure. Put simple, the score is a function of the totals of sentences, words and syllabi - high scores indicate greater readability. We hypothesize that easy to read sentences will be perceived more favorable, yet will also make it more likly to detect BS and down vote it. Therefore, BS should be most effective for lower levels of the Flesch score.

The Model of Reddit Sentiment and BS. One of our central concerns is the causal relationship we hypothesized earlier, i.e., more authentic postings would cause subjects to be more accepting of the posting. The OLS results in Table 1 are an initial assessment of this relationship – but again there is not random assignment here. We just report the correlation between features of a reddit comment, specifically the BS count and Flesch complexity score, the total number of comments in the thread, and the count of net up votes for the positing. Both of our measures have a positive correlation with the count of net up votes on the Reddit posting suggesting that posts with multiple financial 'BS' words get upvoted more more often than those posts that do not contain shallow financial language. Further analysis between the flesch complexity score and the BS score

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.8700	0.0295	131.13	0.0000
bs	0.1365	0.0253	5.41	0.0000
flesch	0.0099	0.0035	2.83	0.0046

support the idea of a negative conditional effects. Instead of using recorded votes

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.8268	0.0304	126.06	0.0000
freq	0.0004	0.0002	1.97	0.0488
flesch	0.0023	0.0036	0.63	0.5286
bs	0.0811	0.0376	2.16	0.0310
freq:flesch	0.0007	0.0000	20.78	0.0000
freq:bs	0.0095	0.0007	14.63	0.0000
flesch:bs	-0.0061	0.0028	-2.21	0.0269
freq:flesch:bs	-0.0002	0.0001	-2.49	0.0129

[2]Flesch R (1948). "A new readability yardstick". Journal of Applied Psychology.

we also model the sentiment[3] of comments for a given post. Again, we use the previous BS dictionary to discern the count of BS words in a submission. As it can be seen the general relationship persists; higher BS scores in a post are correlated with higher positive sentiment in the comments relating to this post.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.0116	0.0077	392.13	0.0000
bs	0.1177	0.0027	43.38	0.0000
flesch.x	0.0040	0.0011	3.71	0.0002
freq	-0.0002	0.0000	-19.07	0.0000

Causal identification

We can use variation in real world financial data as a natural experiment. For one, we may look for examples of shocks to the real economy or to real financial markets that unambiguously reveal fake financial news. Especially salient events in the stock market in 2007/2008 and fluctuations in government bond and annuity rates allow us to relate the sentiment on investment posts marked as fake or bullshit to real market data. And of course we are able to compare these reactions to "control" postings that were unambiguously not face or bullshit. Exploiting similar shocks, we could also observe how external events affect trust in markets and how this loss in trust in turn affects their heuristics.

Since there are several potential "mechanisms" at work here – these could be the message, the participants or the context, for example, we can also use matching to better estimate the "treatment effects." We could, for example, closely match on comments, using the initial post triggering the comment as the treatment variable.

Experimental Interventions

Our causal identification strategy includes an experimental phase in which random assignment of Reddit postings is implemented with the aim of identifying

[3]Based on the Lexicoder dictionary 2015, developed by Mark Daku, Stuart Soroka and Lori Young

those elements of fake news and bullshit that are most effective at moving social media opinion but also those elements that are least effective with the social media community. Using the crowd-sourced coding strategy and trained automated text analysis algorithms described earlier, we expect to be able to categorize our corpus of Reddit postings into the 4 X 5 X 2 matrix in the following table, i.e., classify postings into one of two treatments within each of the 20 cells. Our goal is a minimum of 100 postings for each cell (each cell includes two treatments of 50 postings each) which would give us a total of 2,000 outcomes to analyze.

Table 3: Reddit Postings Treatment Matrix

Bullshit Content	Authority	Emotion	Complexity	Sophistication	Topic
Zero	50/50	50/50	50/50	50/50	50/50
Low	50/50	50/50	50/50	50/50	50/50
Medium	50/50	50/50	50/50	50/50	50/50
High	50/50	50/50	50/50	50/50	50/50

Over the course of about 12 months we will randomly assigning these 2000 postings to three different Reddit financial community threads. On any one day, each Reddit thread would be assigned one posting in the morning and another one in the late afternoon. The treatment effect would be the Reddit metrix discussed earlier: 1) automated analysis of comments; 2) up or down vote; and 3) affect on account Karma. This 4 X 5 X 2 factorial design would allow us to estimate any treatment effects with considerable precision.

The experiment will probe interventions into reducing and potentially critically disapproving, fake news and bullshit financial advice. We cover two main channels to advance the understanding of bullshit reception.

Conclusion and Discussion

Fake news and bullshit are not novel phenomena. There are numerous historical examples of a proliferation of fake news aimed at generating rents or political

advantage for their authors. A particularly interesting example is the paper by Geissler Mesevage (2017). And clearly the Internet revolution has resulted in an explosion of fake news and bullshit in social media.

We are not particularly interested in the proliferation of bullshit. We are interested though in causal claims that exposure to fake news and bullshit affects consumer choices. In particular, we are interested in the effect of this exposure on important financial decisions such as retirement investment decisions. Our contention is that there is a dearth of rigorous causal evidence of such an effect.

This essay describes our bullshit research project that is designed to address this shortcoming and reports on some initial results. Ultimately, our goal is to implement a field experiment design in which individuals interact with their typical menu of social media information in a natural online environment. This is not the case by the way with much of the survey experiment designs that purport to identify the causal effect of fake news or bullshit.

Our online field experiment has both an observational data component and a true experimental component in which subjects are randomly assigned to treatments and control. We report on some initial findings from the first component and sketch out the design features of the second component.

The observational data component builds on the complete corpus of Reddit postings that we have assembled for the period 2008-2015. We have identified those postings that are associated with finance-related sub-Reddits. Ultimately we will score these postings on a series of bullshit and fake news metrics employing crowd-sourced workers but also specialists from business and economic programmes. In our initial phase described here, the crowd-sourced workers rated the profoundness of the post (whether it appears to be insightful and knowledgeable) and whether the post is intelligible. Postings are scored as bullshit if they are high on impressiveness but low on being understandable. We also generated Flesch complexity/readability scores for the posting. These initial metric generation exercises allowed us to gain some insight into what features of Reddit postings correlate with the up or down vote by Reddit. Clearly redditors behave differently depending on the presence of 'BS' and whether this 'BS' is domain specific on financial decision making. Financial decision making specific 'BS' received robust upvotes, even when controlling for the number of comments. An open question

however is how the receptivity to BS changes with the domain. Is financial BS rated differently than generic BS?

These initial results are very encouraging – they clearly raise the possibility at least that the bullshit content of social media posting "cause" individuals to make poor financial decisions. We believe the field experimental phase of the project will provide quite robust evidence of whether there is a bullshit causal effect and its rough magnitude.

We are proposing a unique design in order to accomplish this. In an initial phase we will be using crowd-sourced workers, experts with a business and economic academic background, and automated text analysis to provide a rich detail profile of the bullshit characteristics of Reddit postings. This rich categorization of the Reddit postings will serve as the basis for random assignment of subscribers to financial sub-Reddits to various treatments and controls. Reactions to these random assigned postings will serve as the basis for estimating treatment effects and drawing rigorous conclusions about the causal effect of bullshit on financial decision making.

References

Allcott, Hunt and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. Working Paper 23089 National Bureau of Economic Research.

Ameriks, John, Andrew Caplin and John Leahy. 2003. "Wealth Accumulation and the Propensity to Plan." *The Quarterly Journal of Economics* 118(3):1007–1047.

Berinsky, Adam J. 2017. "Rumors and Health Care Reform: Experiments in Political Misinformation." *British Journal of Political Science* 47(2):241?262.

Calvet, Laurent E., John Y. Campbell and Paolo Sodini. 2009. "Measuring the Financial Sophistication of Households." *The American Economic Review* 99(2):393–398.
URL: *http://www.jstor.org/stable/25592430*

Duflo, Esther and Emmanuel Saez. 2003. "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment." *Quarterly Journal of Economics* 18:815–842.

Geissler Mesevage, Gabriel. 2017. "Fraud and Financial Manias: Estimating the Number of Bubble Companies during the Railway Mania of the 1840s.".

Hastings, Justine S., Brigitte C. Madrian and William L. Skimmyhorn. 2013. "Financial Literacy, Financial Education, and Economic Outcomes." *Annual Review of Economics* 5(1):347–373.

Mani, Anandi, Sendhil Mullainathan, Eldar Shafir and Jiaying Zhao. 2013. "Poverty Impedes Cognitive Function." *Science* 341(6149):976–980.

Pennycook, Gordon and David G. Rand. 2017. "Who Fall for Fake News: The Roles of Analytic Thinking, Motivated Reasoning, Political Ideology, and Bullshit Receptivity.".

Pennycook, Gordon, Tyrone D. Cannon and David G. Rand. 2017. "Implausibility and illusory truth: Prior exposure increases perceived accuracy of fake news but has no effect on entirely implausible statements.".

Thaler, Richard H. and Cass R. Sunstein. 2009. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Penguin.

Tversky, Amos and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185:1124–1131.