

Causal Analysis of Online Outcomes of Interactions between Organic and Inorganic Accounts

Onur Savas, Benjamin A. Ortiz, and Ian McCulloh

Accenture, Washington, DC, 2005 USA

{onur.savas,benjamin.a.ortiz,ian.mcculloh}@accenturefederal.com

Abstract. Despite great advancements in the understanding of propagation and evolution of disinformation and the inorganic accounts involved in this process, little is known regarding whether and how much interactions between organic and inorganic accounts causes changes in online outcomes. Hence, we pose the question: “What are the most likely explanations for changes in online outcomes of organic users (e.g., ordinary social media users) after interacting with inorganic accounts (e.g., social bots, cyborgs, trolls, sock puppets)?” To answer, we provide a framework to quantify the causal effect of these interactions on online outcomes. As part of this framework, we first formulate the question as a causal inferencing problem by adopting a Neyman-Rubin causal model (NRCM), incorporate data onto this model, and apply statistical matching techniques to quantify Average Treatment Effect (ATE). This framework enables causal inference from observational data and provides an inexpensive alternative to randomized controlled trials (RCTs), which can be impermissible, infeasible, or not scalable in novel information environments. Using a dataset of tweets that covers more than 16 weeks, we show that the interactions account for 7.4% of the change in online sentiment and emotion change, while the remaining 92.6% can be explained by other covariates such as gender and ideology. We also observe that changes in anger, fear, and sadness are more likely to be explained (more than 12%) by interactions between organic and inorganic accounts.

Keywords: Social media, social bots, cyborgs, trolls, sock puppets, online behavior change, Neyman-Rubin model, causal inferencing.

1 Introduction

Social media and networking have become a crucial part of our democracy by lowering the barrier for citizens to participate in public discussions, enabling real-time news and information discovery, rebuilding trust in public institutions with organic engagements, and informing the public about policy changes and its benefits. These low barriers, however, enabled the manipulation of public discourse by adversarial actors, undermining our democracy and its institutions.

We now understand the intent and actions of these adversarial actors better. For example, these adversarial actors created and nurtured inorganic accounts such as social bots, cyborgs, trolls, and sock puppets to interfere with our democracy by fabricating

information that disseminates deceptive content, or grossly distorting actual news reports to affect public discourse and elections [4][8][11][32]. Recent research has made great advancement towards identifying these inorganic accounts [6][9][11][14][33][35], understanding how they coordinate [1][6][12][24][25], analysis of the content of their messages [8][21][24][26], and target audiences involved [4][5][19]. However, little is known regarding *whether* and *how much* the interactions between organic and inorganic accounts cause changes in online outcomes.

1.1 Introduction of Terminology

Before proceeding further, we introduce basic concepts and terminology regarding our approach: A *unit* is an entity upon which a treatment will operate, at a particular time. A *treatment* is an intervention, the effects of which (on some particular measurement on the units) we wish to assess relative to no intervention (i.e., the control). *Potential Outcomes* are the values of a unit’s measurement of interest after (a) application of the treatment and (b) non-application of the treatment (i.e., under control). For each unit, *causal effect* is the comparison of the potential outcome under treatment and the potential outcome under control. *Selection Bias* arises from the fact that treated individuals differ from the non-treated for reasons other than treatment status per se. *The Fundamental Problem of Causal Inference* is such that we can observe at most one of the potential outcomes for each unit.

The goal of causal inference is then to find the causal effect(s) by overcoming the challenges introduced by the fundamental problem of causal inference.

1.2 Contribution

In this study, we are interested in answering “*What are the most likely explanations for changes in online outcomes of organic users (e.g., ordinary social media users) after interacting with inorganic accounts (e.g., social bots, cyborgs, trolls, sock puppets)?*” Ideally, a randomized controlled trial (RCT) is best suited to answer this and similar questions since it is the gold standard of understanding the effectiveness of a new intervention or treatment [13]. By randomization, it is possible to reduce bias and provide systematic study to examine cause-effect relationships between an intervention and outcome. By balancing participant characteristics (both observed and unobserved) between the groups, we allow attribution of any differences in outcome to the study intervention. This is infeasible with any other study design.

To design an RCT for our setting, we need, at minimum, to create and nurture the inorganic accounts, curate their content, and control who the content is exposed to / interacted with. We also need a measure of effectiveness in terms of a quantifiable online outcome. However, RCTs of this kind is generally neither *permissible* (e.g., due to restrictions by social media platforms) nor *feasible* (e.g., due to the infeasibility of getting informed consents from a very large population) nor *scalable*¹. To provide a permissible, inexpensive, and scalable alternative to RCTs, we propose a causal inferring framework. However, we face a number of challenges in formulating the problem, which will address our research goals.

First, we need to focus our attention on inorganic accounts, and develop an appropriate dataset with the right attributes that will enable us to pursue our research objective and enable causal analysis. Once we have a list of inorganic accounts, we can

study the organic accounts that already do follow the accounts. These subjects are going to be the collection of social media users whose treatment effects we are going to analyze. Their changes in online outcomes (e.g., behavior/ belief) are the effects that we are going to measure. Some of these will have engaged in interactive discussions with inorganic accounts, and some will have not. This distinction will ultimately divide our users into two groups on which we can base our observational study.

Second, for all those organic accounts, we need to have a measure of online outcome (belief) about that the inorganic account is attempting to affect. This is especially challenging because it requires us to identify all the potential subjects that inorganic accounts are discussing with organic ones and determine the sentiment of a particular human toward a particular topic. We address this challenge by focusing on a subset of manually curated content, for which we know the intent.

Third, we need to have statistical methods to determine whether changes in online outcomes (belief) can be attributed to bot/human interaction. This is a necessary for the purpose of evaluating the intervention of bots in human discourse by isolating the cause and removing other confounders from the explanation of differences. We address this by adopting the Neyman-Rubin Causal Model (NRCM) framework [26][29] which views causal effect as comparisons between potential outcomes defined on the same units.

The proposed framework requires neither the interactions between organic and inorganic accounts nor the factors or outcomes to be under control. We collect data from social media platforms driven by the model’s requirements, then enrich and incorporate this data into the model to solve for Average Treatment Effect (ATE), which quantifies the magnitude of causal effect of any treatment.

In the rest of this paper, we first introduce the causal model, adopt it for our setting, discuss practical consideration, and present our numerical findings. As appropriate, we give representative examples of organic and inorganic accounts, online behaviors of consideration, and interactions. We conclude with future directions and discuss limitations of the study.

2 Study Design

2.1 Neyman-Rubin Causal Model (NRCM)

NRCM is a basic statistical model of cause and effect based on the idea of potential outcomes. The Average Treatment Effect (τ_{ATE}) is the quantity of interest. Under general NRCM, we are given a table $R(T, X, Y(0), Y(1))$, shown in Table 1, with N rows called units U_i , indexed by $i = 1, 2, \dots, N$. The variable T is a binary outcome whether or not the unit was treated. X is a vector of K covariates which are attributes of each unit unaffected by the treatment T . $Y(0)$ and $Y(1)$ are two attributes known as the potential outcomes. $Y(0)$ is the outcome if the unit was exposed to treatment and $Y(1)$ is the outcome if the unit was exposed to control. The Treatment Effect for U_i is defined as $T_i = Y_i(1) - Y_i(0)$.

Table 1. The table $R(T, X, Y(0), Y(1))$.

U	T	$X = [X_1, X_2, \dots, X_K]$	$Y(0)$	$Y(1)$	$Y(1) - Y(0)$
1	T_1	$X_1 = [X_{11}, X_{12}, \dots, X_{1K}]$	$Y_1(0)$	$Y_1(1)$	$Y_1(1) - Y_1(0)$
2	T_2	$X_2 = [X_{21}, X_{22}, \dots, X_{2K}]$	$Y_2(0)$	$Y_2(1)$	$Y_2(1) - Y_2(0)$
...
N	T_N	$X_N = [X_{N1}, X_{N2}, \dots, X_{NK}]$	$Y_N(0)$	$Y_N(1)$	$Y_N(1) - Y_N(0)$

The goal of causal analysis is to compute the Average Treatment Effect, namely $\tau_{ATE} = \frac{1}{N} \sum_{i=1}^N Y_i(1) - Y_i(0)$. The underlying difficulty is that, in reality, we only can observe one outcome for each unit i : A unit is either treated or not. The other missing value prevents us from computing τ_{ATE} using and is known as the fundamental problem of causal inference. Therefore, in order to compute τ_{ATE} , further assumptions such as matching or blocked randomization are frequently needed [26][29].

2.2 Adoption of the Model

To propose our problem under this framework, we need to be able to well-define the following variables as it relates to our current problem.

U_i : An individual organic user/account indexed by $i = 1, 2, \dots, N$.

T_i (Treatment Assignment): 0 or 1 variable indicating whether user i was treated or not. By treated, we mean that the organic user had measurable interaction with inorganic accounts. For example, by reposting, liking or favoriting certain posts of a troll.

Y_i : Online outcome change measured in U_i . For example, organic user i could have exhibited a change in language after interacting with inorganic accounts. This would mean that the inorganic account is potentially exerting some sort of an impact on organic accounts.

$X_i = [X_{i1}, X_{i2}, \dots, X_{iK}]$: List of K attributes or confounding variables about each organic account i . Ideally, we want to include demographics such as gender, age, location, and income and education. We also want to include previous beliefs and political stances.

In our solution, the data to be modeled and input to causal analysis is prepared as a structured in Table 2 as below with representative covariates (e.g., age, gender).

Table 2. The adopted model's table $R(T, X, Y(0), Y(1))$.

U (User handle)	T (Interacted w/ Inorganic Account)	X_1 (Age)	X_2 (Gender)	X_3 (Location)	...	Y (Outcome Change)
@abc	1	X_{11}	X_{12}	X_{13}		Y_1
@def	0	X_{21}	X_{22}	X_{23}		Y_2

@ghi	0	X_{31}	X_{32}	X_{33}		Y_3
...
@xyz	1	X_{N1}	X_{N2}	X_{N3}		Y_N

The principal problem in the estimation of treatment effects is **selection bias**, which arises from the fact that treated individuals differ from the non-treated for reasons other than treatment status per se. To account for the lack of gold standard in nonrandomized control trials, we compute ATE using a balancing function to group units with similar covariates together. This is what one approach known as matching tries to solve.

3 Practical Considerations

In this section, we propose strategies to find both organic and inorganic accounts (hence find units U), assign treated / control groups (units for which $T_i = 0$ or 1), quantify online outcomes (Y_i), and infer covariates (X).

3.1 Identification of Inorganic Accounts

The first step is to identify inorganic accounts by distinguishing it from organic accounts. The methods and algorithms in [6][9][11][14][33][35] can be leveraged to find inorganic accounts. Most of the literature have focused on finding social bots, where factors such as regular interposting times, posting rate, unbalanced following/follower ratio have been used as distinguishing factors. However, we focus our attention not only to social bots but other types of inorganic accounts as listed below.

Troll: A social media user that harasses, abuses, and attempts to provoke/trigger emotions in other social media users.

Sybil: A known impostor or fake account, such as an impostor account of a real celebrity, sometimes for parody.

Cyborg: A human user that leverages third party apps to schedule, amplify, and to generally enhance their capabilities beyond the limitations of a normal human.

Sock puppet: An online identity used for purposes of deception.

Revenant: A subcategory of cyborgs that respawns a new account with similar content when suspended by the social media platform.

Influence Bot: A social media account that illustrates an intent to persuade and to perturb information propagation. Influence bots can often be found working in groups and can be both fully and partially automated.

To find inorganic accounts with an intent to influence, psycholinguistic models to detect biased language [28] or persuasive language can be used [16]. To detect sock puppets, key features that distinguish sock puppets from humans including but not limited to amount and origin of original text and image posts, pattern of life, social graph, evolution and consistency of account (repurposing) can be used. As many sock

puppets are managed by a single operator, account similarity may also discriminate between the three classes.

3.2 Assessing Online Outcomes

To compute Y , we need to find online outcome change. We focus on three types of online outcome change: (i) use of discursive/linguistic patterns, (ii) sentiment towards a narrative/topic, and (ii) online pattern of life (e.g., interaction with its social network).

To measure *discursive/linguistic patterns*, we will measure demonstrable discourse change by users online. For example, as demonstrated in [26], to identify the radicalization and recruitment of Twitter and blog readers to ISIS, the researchers used the change in tone and in topic after exposure to ISIS propaganda. In another example, [36] identified the shift from singular personal pronoun (“I”) to plural personal pronoun (“we”) as a transition point where an individual finds his/her identity to be closer to the collective, as a direct outcome of online campaigns.

The *sentiment against a topic* can similarly be measured by first identifying the topic/narrative and then applying sentiment analysis. In fact, the USASOC C-ISIS simulation focusing on cognitive aspects of IO narratives [21] demonstrated this as a measurable outcome of the IO messaging.

To measure *online pattern of life*, we can first find online indicators of patterns of life that include but are not limited to who the account interacts with, the groups she/he subscribes to, the events she/he attends, his/her friends/followers. We also hypothesize that, due to homophily, the variation in his/her network will be an indicator of the account itself (e.g., friending with radicalized accounts might be an indicator that the account itself is being radicalized).

3.3 Inferring Covariates (Alternate Explanations)

An important aspect of causal inferencing is to account for covariates or confounding factors. While it is impossible to account for every confounding factors, studies such as [3] show demographics and psychographics can be important factors in how online users are influenced or not. Hence, we start with inferring demographics and psychographics of the organic accounts. In brief, one can use age, gender, income, occupation, ethnicity, highest level of education, religion, location, and political affiliation. Many social platforms do not provide, these parameters need to be estimated. In Table 3, we provide algorithms to estimate these demographics, which can then be used as covariates for causal inferencing.

Table 3. Inference of covariates.

Covariate Name	Method of Estimation
Age	Given a first name, estimate the probability that they were born in each year between 1920 – present day [14].
	Exploitation of age-related words or phrases in microblogs [21].
Gender	Searches for a pronoun announcement in each of fol-

	lowers' profiles [9]. Employing Support Vector Machine (SVM) to estimate bloggers' gender from blog posts [17].
Income / Occupation	Analysis of Twitter behavior [26].
Ethnicity	Correlating the last name with data from the U.S. 2000 Census [19].
Location	A review of approaches for location estimation in Twitter [37]
Political Ideology	The structure of the users' social networks can be a source of information about their ideological positions [6].

4 Numerical Findings

We have collected more than 40k number of tweets from an inorganic account (a troll) between January 7, 2019 and May 7, 2019 (approximately 16 weeks). We have identified the most interacted tweets, which contain deceptive content. A representative such tweet is shown in Fig 1.



Fig. 1. Representative tweet from the inorganic user.

We have then identified 10 organic accounts, which retweeted or replied to one of the previously identified tweets. This set of organic users constitute the *treated group*, i.e., for which $T_i = 1$. We also identified 10 organic accounts, which followed this troll but did not interact. This set of organic users constitute the control group, i.e., for which $T_i = 0$. Overall, we have $N = 20$ units.

We have then pulled timeline of the treated group's tweets, identified the time when these users interacted with the troll, and separated the tweets into two bins: *before* and *after* the interaction. We have done the same for the control group. However, since the control group never interacted with the troll, we have separated the tweets at date/time of the troll's original tweet.

We then identified the following set of online outcomes as the emotional changes (note the superscript such that it does not confuse with unit number).

$Y^1 .. Y^{10}$: 10 emotion/sentiment measurements, namely anger, anticipation, disgust, fear, joy, negative, positive, sadness, surprise, trust.

We have then computed the following covariates, which are binary as indicated.

X_1 : Gender (Male or Female)

X_2 : Political Ideology (Liberal or Conservative).

Using these computed values, we have then run the NRCM model using the Coarsened Exact Matching (CEM) R package. The results are presented in Table 4.

Table 4. Explanation of interactions on online outcomes.

Variable	Online Outcome	ATE Explained by Interactions
Y^1	Anger	14.5% ($p = 0.22$)
Y^2	Anticipation	0.7% ($p=0.32$)
Y^3	Disgust	0.9% ($p=0.13$)
Y^4	Fear	23.7% ($p=0.15$)
Y^5	Joy	3.3% ($p=0.12$)
Y^6	Negative	5.0% ($p=0.25$)
Y^7	Positive	4.3% ($p=0.17$)
Y^8	Sadness	12.3% ($p=0.21$)
Y^9	Surprise	7.6% ($p=0.22$)
Y^{10}	Trust	1.5% ($p=0.31$)

The third column in the above table shows the percentage of online interactions explaining the outcome change (the variable in the first column). Overall, the interactions account for 7.4% of the change in online sentiment and emotion change, while the remaining 92.6% can be explained by other covariates such as age or ideology. We also observe that changes in anger, fear, and sadness are more likely to be explained (more than 12%) by interactions between organic and inorganic accounts.

5 Conclusion and Discussions

The major contribution of this study is a computational framework that explains the changes in online outcomes arising from interactions of organic users with inorganic users. In particular, we show how this problem can be mapped into a Neyman-Rubin causal model. We review practical considerations including but not limited to inference of parameters in the model. We show, in a use case, that interactions between organic and inorganic accounts account for more than 23.0% of change in fear. Emotions such as sadness and anger can also be explained by more than 10% by interactions.

The major limitation of this ongoing study arises from the lack of sufficient data as can be evidenced by high p values. We envision more statistically significant results can be obtained by at least using thousands of online accounts. Exploitation of only 2 covariates (i.e., alternative explanations) is another limiting factor. While it is impossible to include every possible explanation, we will add more demographics, psychographics as well as metrics that capture the users' network in a future study.

References

- [1] N. Agarwal et al. *Examining the Use of Botnets and their Evolution in Propaganda Dissemination*. NATO Strategic Communications Centre of Excellence, volume 2, 2017.
- [2] T. Althoff, P. Jindal, and J. Leskovec, "Online Actions with Offline Impact: How Online Social Networks Influence Online and Offline User Behavior," in Proc. of the ACM WSDM 2017.
- [3] S. Aral, and D. Walker. *Identifying Influential and Susceptible Members of Social Networks*. Science, 337(6092):337-341, 2012.
- [4] A. Badawy et al. *Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign*. 2018.
- [5] C. A. Bail et al., *Exposure to Opposing Views on Social Media can Increase Political Polarization*. PNAS September 11, 2018 115 (37) 9216-9221; first published August 28, 2018.
- [6] P. Barbera. *Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data*. Political Analysis, Volume 23, Issue 1, Winter 2015 , pp. 76-91.
- [7] D. Beskow, and K. M. Carley, "Introducing Bothunter: A tiered Approach to Detection and Characterizing Automated Activity on Twitter," in Proceedings of the 2018 SBP-BRiMS.
- [8] A. Bovet, and H. A. Makse. *Influence of Fake News in Twitter During the 2016 US Presidential Election*. Nature Communications 10, Article number: 7 (2019).
- [9] A. J. J. Davis. 72% Of The People I Follow On Twitter Are Men. <https://emptysqua.re/blog/gender-of-twitter-users-i-follow/> (accessed May 15, 2019).
- [10] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. *The Rise of Social Bots*. Communications of the ACM, July 2016, Vol. 59 No. 7, Pages 96-104.
- [11] P. Fornacciari, M. Mordonini, A. Poggi, L. Sani, and M. Tomaiuolo. *A Holistic System for Troll Detection on Twitter*. Computers in Human Behavior, Volume 89, December 2018, Pages 258-268.
- [12] R. Galeano et al.. *Botnet Evolution During Modern Day Large Scale Combat Operations*. In "Information Operations: Perceptions are Reality", Chapter 10, 2018.
- [13] E. Hariton, and J. J. Locascio. *Randomised controlled trials—the gold standard for effectiveness research*. BJOG Research Methods Guides, 2018.
- [14] How to Estimate Someone's Age, <https://logicalerrors.wordpress.com/2015/08/16/how-to-estimate-someones-age-based-on-their-twitter-username-using-t/> (accessed May 14, 2019).
- [15] T. Khaund et al., "Analyzing Social and Communication Network Structures of Social Bots and Humans," in Proc. of the 2018 ASONAM. ,
- [16] T. Khazaei X. Lu, and R. Mercer, "Writing to Persuade: Analysis and Detection of Persuasive Discourse," In iConference 2017 Proceedings (pp. 203–215), 2017.
- [17] D. Kobayashi, N. Matsumura, and M. Ishizuka, "Automatic Estimation of Bloggers' Gender," in Proc. of the ICWSM'16, Boulder, CO USA, 2016.
- [18] A. D. I. Krame et al. *Experimental Evidence of Massive-scale Emotional Contagion through Social Networks*. PNAS, June 17, 2014, 111 (24) 8788-8790; first published June 2, 2014. <https://doi.org/10.1073/pnas.1320040111>.
- [19] A. Mislove et al., "Understanding the Demographics of Twitter Users," in Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, 2011.
- [20] B. M. Monsted et al. *Evidence of Complex Contagion of Information in Social Media: An Experiment using Twitter Bots*. PLoS ONE 12(9): e0184148, 2017. <https://doi.org/10.1371/journal.pone.0184148>

- [21] C. Moore et al., “Counter-Da’esh Influence Operations: Cognitive Space Narrative Simulation Insights,” May 2016.
- [22] N. Moseley. *Using Word and Phrase Abbreviation Patterns to Extract Age from Twitter Microtexts*. MS Thesis, Rochester Institute of Technology, 2013.
- [23] B. Nimmo. *Measuring Traffic Manipulation on Twitter*. Working Paper 2019.1. Oxford, UK: Project on Computational Propaganda. comprop.oii.ox.ac.uk.
- [24] B. Nimmo et al. *#TrollTracker: Twitter Troll Farm Archives*. <https://medium.com/dfrlab/trolltracker-twitter-troll-farm-archives-8d5dd61c486b> (accessed May 13, 2019).
- [25] A. Olteanu, O. Varol, and E. Kiciman, “Distilling the Outcomes of Personal Experiences: A Propensity-Scored Analysis of Social Media,” in Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, 2017.
- [26] J. Pearl. *Causal inference in statistics: An overview*. UCLA Computer Science Department, Technical Report R-350, 2009.
- [27] D. Preoțiuc-Pietro, S. Volkova, V. Lampos, Y. Bachrach, and N. Aletras. Studying User Income through Language, Behaviour and Affect in Social Media. PLoS ONE, September 22, 2015. <https://doi.org/10.1371/journal.pone.0138717>
- [28] M. Recasens, and D. Jurafsky, “Linguistic Models for Analyzing and Detecting Biased Language,” in Proc. of the ACL, December 2012.
- [29] D. Rubin. Causal Inference Using Potential Outcomes. J. Amer. Statist. Assoc., 2005.
- [30] H. Saif et al., “A Semantic Graph-Based Approach for Radicalisation Detection on Social Media,” in Proc. of the European Semantic Web Conference (ESWC 2017), pp. 571-587, 2017.
- [31] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer. *The Spread of Low-credibility Content by Social Bots*. Nature Communications, 9:4787, 2018.
- [32] M. Stella, M. Cristoforetti, and M. De Demonico, “Influence of Augmented Humans in Online Interactions during Voting Events,” arXiv preprint arXiv:1803.08086 (2018).
- [33] V. S. Subrahmanian et al. 2016. *The DARPA Twitter Bot Challenge*. Computer 49, 6 (June 2016).
- [34] S. J. Taylor, and D. Eckles. *Randomized Experiments to Detect and Estimate Social Influence in Networks*. Complex Spreading Phenomena in Social Systems, pp. 289-322, 2018.
- [35] O. Varol et al., “Online Human-Bot Interactions: Detection, Estimation, and Characterization,” in Proc. of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017).
- [36] S. T. Yuce et al. *Bridging Women Rights Networks: Analyzing Interconnected Networks in Online Collective Actions*. Journal of Global Information Management 22(4):1-20, September 2014.
- [37] X. Zheng et al. *A Survey of Location Prediction on Twitter*. IEEE Transactions on Knowledge and Data Engineering, Volume: 30, Issue: 9, September 2018.

¹ For example, in [19], a botnet of 39 human-like inorganic accounts have been created to observe information contagion dynamics. Other large-scale experiments in social media, though not in the context of organic-inorganic account interactions, have also been conducted [2][3][16][34].