Idiomatic Application of Causal Analysis to Social Media Timelines: Opportunities and Challenges

Golnoosh Farnadi
Ghent University & Katholieke Universiteit Leuven
golnoosh.farnadi@ugent.be

Emre Kiciman
Microsoft Research
emrek@microsoft.com

Causal inference, from experimental and observational studies, is critical to answering important questions in natural, social and digital systems. Unfortunately, applying causal inference to large systems—such as markets, societies or even teams of people—presents critical challenges in causal inference due to network effects, feedback loops and other complications. While many causal methods have been introduced and are applicable to some of these problems, their use requires careful thought and adaptation by experts. But *what if* we could identify a (large) class of important questions that could be answered without repeated expert intervention? We identify such a broad class of simple questions about individual experiences—essentially, what happens after a person takes some action or has some experience—that can be answered through analysis of a large-scale corpus of individual-level social media timelines under ignorability and SUTVA assumptions. Our goal is to create a framework for data processing and causal inference methods that can best answer these action-outcome questions from social media timelines.

Providing answers for this class of causal questions using our simplified causal framework is of interest to individuals, scientists and policy makers. For example, individuals may look for ways to better understand the consequences of their decisions. Using our framework, they can effectively aggregate the experiences of hundreds of millions of people, many of whom have made similar decisions and reported on their experiences. The inferred causal outcomes can help people make better decisions, from selecting better products to making better career and life decisions. For scientists and policy makers, understanding various situations and their possible implications of taking actions provides an opportunity to better understand phenomena of social importance, e.g., bullying, planning for retirement, college graduation and unemployment, among many others. Advantages in using social media data for this purpose are as follows. First, results are grounded based on the real experiences of people who have taken an action which increases the reliability of the results. Second, while some goals are common and there are many web articles and advice about them, using social media platform increase the chance of finding an answer. And third, given the preponderance of data, we may provide personalized answers tailored to the asker.

We define our causal framework as follows: let T be the set of experiences (i.e., treatments) we wish to consider and X the set of users. Each user x is characterized by a vector of covariates (e.g., textual features extracted from their posts) $x \in \mathbb{R}^n$. We are interested in the case of binary set of experiences, i.e., $T = \{0, 1\}$, e.g., $T = 1$ is taking a certain medicine. Therefore, for any given T, we find all social media users who have reported in their posts that they had that experience (i.e., took the medicine); The set of users who have the experience i.e., $T = 1$ is often known as "treated" group in causal inference literature and the set of users who do not have the experience, i.e., $T = 0$, is known as the "control" group. Let Y be the set of possible outcomes. We consider both X and Y to be represented by textual features. To identify Y, we analyze all posts come after the one containing T from the time-lines of the treated group. To represent X and Y of the control group, we randomly select a time-stamp ts from the set of timestamps of the treatment posts and then split the time-lines of the control group to posts come before and after the selected time-stamp ts. Any given $y \in Y$ is either y_1 or y_0, where y_1 means a textual feature y occurs and y_0 means a textual feature y does not occur. Note that we can only observe one of the outcomes for each user x. There is a large literature on various approaches to deal with estimating average causal effects (i.e., $ATE_{x \sim D(X)} = \mathbb{E}[Y_1(x) - Y_0(x)]$). Part of this work which employs propensity score analysis has been published in (1).

To identify the best inference methods for our scenario, we empirically evaluated several standard approaches, using both naturalistic data as well as synthetic and mixed naturalistic and synthetic datasets. We investigate four categories of these approaches, namely 1- matching (e.g., propensity score and Mahalanobis), 2- weighting (e.g., inverse propensity score), 3- regression adjustment, and 4- doubly robust methods. We gathered three months Twitter data with more than 69M tweets of

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

50k users to evaluate the framework. We select a diverse set of 39 experiences on various domains including disease, pharmacy, society issues, finance, and business. We first compare the performance of each algorithm based on precision@10 of ATE scores using the annotations of the crowd-source workers. Most of the algorithms perform well with 60-100% accuracy using naturalistic data. While many associations between the outcome and the treatment are extracted by the causal inference algorithms, labelling those outcomes as causal require domain knowledge and therefore annotating the outcomes by crowd-source workers is a challenging task. For instance, we find that users posting about taking *Xanax* in their tweets, talk about smoking and drinking in their post-treatment posts. This behaviour has been recognized by correlation-based, matching and covariate adjustment methods while outcomes which indicate the side effects such as puking and sleep disorder are recognized by doubly robust method. Weighting methods work similar to the matching methods in recognizing irritability but also recognize category of the similar drugs such as *klinopin* and *benzos* which stands for *benzodiazepines*.

To investigate the differences among algorithms in our dataset and remove human judgment, we design experiments with naturalistic data with injected synthetic ground-truth outcomes. We study various parameters according to the underlying data: the size of the treated and control groups, the portion of the outcome among the treated and control groups, dependency among treatment and covariates, and dependency among outcome and covariates. We observe that the behaviour of the inverse propensity score weighting approaches depends on the portion of the outcome among treated and control groups, and the weights may be inaccurate or unstable for users with a very low probability of receiving the treatment. The behaviour of the matching techniques are similar to the correlation-based methods and by removing the dependency between the treatment and covariates, weighting methods become stable and their performance get closer to the correlation-based methods. Note that this behaviour is not always desirable, e.g., "Simpson's paradox". Depending upon the causal question, we examine the critical factors to automatically choose a suitable technique which enhances both reliability and validity of the outcome. This analysis indicates the potential to develop a framework that best answer causal questions from social media timelines without causal inference expertise.

We borrow concepts from the causal inference literature, however it is important to note that our framework cannot satisfy all the key causal assumptions. For measuring causality in social media data, we consider *ignorability* which assumes there is no unmeasured cofounders, i.e., $T \perp\!\!\!\perp (Y_0(x), Y_1(x)|x)$. However, social media data may not fully satisfy ignorability as social intervention may happen. Many of the estimation issues raised by social interventions are discussed in Baird et al (2). Another strong assumption in causal inference problems is the *balance* assumption. Balance means the distributions of relevant pre-treatment variables should not differ for the treatment and control groups,i.e., $\mathbb{P}(T = t|X = x) > 0, \forall t, x.$, However, in high-dimension setting, where the number of covariates (features) n are large, i.e., $n >> 0$, such as our setting, it is often impossible to guarantee the balance assumption. And as discussed by Athey et al (3), complete balancing of all covariate is not always necessarily nor needed. Another challenge that comes with social media textual data is that causal interpretation, which can be influenced by significant bias due to population biases as well as self-reporting biases (4). The absence of written experiences in time-lines of users does not necessarily mean an experience did not happen. We rely on time-lines of users to split covariates from outcomes, however it is important to note that the exact time that certain experiences happen in real life for a user may not match the order that they report the experience in their time-lines.

In this abstract, our aim was to introduce an open-domain framework that separates causal inference expertise from domain knowledge that semantically interprets the results. We study various characteristics of the social media time-lines, which allow us to select a suitable causal method given the underlying properties of the data and causal question. We address the opportunities and challenges, and there are many open questions remain for our future work: How to present the outcomes to the user? How to select a causal method based on different types of causal questions? How to leverage the framework with current quantitative and qualitative methods to understand societal phenomena?

References

[1] A. Olteanu, O. Varol, and E. Kiciman, "Distilling the outcomes of personal experiences: A propensity-scored analysis of social media," in *Proc. of The 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing*, 2017.

[2] S. Baird, J. A. Bohren, C. McIntosh, and B. Ozler, "Designing experiments to measure spillover effects," in *Policy Research Working Paper Series 6824, The World Bank.*, 2014.

[3] S. Athey, G. W. Imbens, and S. Wager, "Efficient inference of average treatment effects in high dimensions via approximate residual balancing," *arXiv preprint arXiv:1604.07125*, 2016.

[4] F. Diaz, M. Gamon, J. Hofman, E. Kiciman, and D. Rothschild, "Online and social media data as a flawed continuous panel survey," tech. rep., Working Paper http://research. microsoft. com/flawedsurvey, 2014.