Detecting Shifts in Propensity Score Stratification when using Relational Classifiers for Network Data

Francesco Marcantoni
University of Illinois at Chicago
Polytechnic University of Turin
fmarca2@uic.edu

Marco Mele
University of Illinois at Chicago
Polytechnic University of Turin
mmele2@uic.edu

ABSTRACT

We address the problem of reviewing the Stratified Propensity Score Analysis used for evaluating the Average Treatment Effect in treatment effect observation when we deal with relational data. It is common practice to stratify the populations of treatment and control groups with Propensity Score analysis to divide the population in *strata* with similar propensity in showing the treatment effect; this approach makes use of general classifiers to learn the Propensity Score Estimation for each unit, missing the benefits that a relational classification can bring when it comes to relational data, when units are linked within and among treatment and control groups.

KEYWORDS

Propensity Score; Average Treatment Effect; Casual Inference

1 INTRODUCTION

The Rosenbaum and Rubin [21] introduces the Propensity Score as

> the conditional probability of assignment to a particular treatment given a vector of observed covariates.

The Propensity Score analysis is used to reduce bias in observational data using stratification on treatment and control group. Given a treatment z, the Propensity Score is the conditional probability that a unit with observed covariates $\mathbf{x}$ will be assigned to the treatment, i.e.

$$e(\mathbf{x}) = \Pr(z = 1|\mathbf{x}) \tag{1}$$

The stratification on the propensity score has been proven to balance the distribution of $\mathbf{x}$ among treatment and control group in each stratum of units that are homogeneous in $e(\mathbf{x})$ [20].

One of the most important problem in randomized trials is to assess the generalizability of the results to a target population [23]. In real world data treatment and control group can have large differences in their observed covariates leading to a biased evaluation of the treatment effect [9].

Consider a generic setting of a randomized trial to estimate the effects of a treatment z on the treatment group $\mathcal{T}$ with respect to a control group C, where units have been randomly assigned to either $\mathcal{T}$ or C. When analyzing the effectiveness of z, we measure the Average Treatment Effect as

$$\text{ATE} = \frac{1}{n}\sum Y_i(1) - Y_i(0) \tag{2}$$

where $Y_i(1) - Y_i(0)$ is the difference in potential outcome for the unit i [23].

In observational, non-randomized studies, the differences in observed covariates can produce uncontrolled bias in a unit membership to either the treatment or the control group [9]. The point in observational studies is that, unlike an experiment, there is no systematic method to maintain a control group such that it is not biased the way the treatment is assigned to units, that can be the result of self-selection or systematic judgement by the researcher [8].

On the other side of the medal, Rubin [22] summarizes some limitations of the propensity score matching, one being the impossibility to control for unobserved selection bias: PS cannot adjust for hidden selection bias [12]: this motivates our intent in including network information in propensity score estimation on relational data learning the estimation with a relational classifier.

There are some ongoing controversy on which variables to include in the propensity score learning, either only those variables that are predictors of the treatment assignment, or all variables potentially related to both the treatment and the outcome, or still only those strictly related to the outcome [2].

In our work we focus on Relational Data coming from Social Media; online user–generated content bases are a huge collection of people's thoughts regarding the most various matters. Not only people use online social networks for communication, but they also intensively disclose short and frequent statements about their experiences, feelings, and situations. A particular aspect that arises from this collection of thoughts is the quest for the tendency, in the history of a person's publications, to find signals of a potential outcome as a consequence of a previously experienced situation.

In this setting, our particular interest is to try and determine whether there is an impact on users' propensity of exposure to a treatment when we learn it including their social interconnections in the covariates, to see if they hide unobserved covariates potentially influencing the estimation. Social networks have different ways of entity relations: Twitter, for instance, defines unidirectional relationships between users in terms of *following/followers*, mentioning, and retweeting. We find of interest to exploit the relational nature of Twitter users in the attempt of determining causality between a set of target outcomes and personal experiences.

People have the need of continuously investigating details of situations they find themselves in, and that possibly do not understand well. Someone experiencing a social, physical or whatsoever situation they are not familiar with, is struggling in looking for knowledge of the possible unfoldings of their condition [19]. This does apply for a variety of domains: from the mining of the possible consequences of specific symptom, to the potential results gained from workout, or the investigation on the process for filing for divorce.

The majority of people consistently follow the crowds—a phenomenon that has been brought to its largest extent in such a strongly connected world we live in: social interconnections enable the spread of information as, among other things, preferences and behaviors [7, 10]. On the other hand, a minority of the population has an interest in *niche content* and, anyway, everyone is a bit eccentric, tending to both popular behaviors and specialty ones [11].

We can translate the concept of non-randomized observational study on treatments and effects into the quest for situations and outcomes of personal experiences; this is not only one person's interest; scientists are always looking for answers to questions of public interest, like: do mental health content in social media influence shifts in one's suicidal propensity [7]? For this purpose, a lot of *knowledge bases* and services provide people a pool of information that may help understand some common situation domain, to help them in their processes of decision making or goal achievement [19].

2 FORMAL PROBLEM DESCRIPTION

Propensity Score Stratification is largely used to address to problem of reducing bias when assessing causality in observational data to measure the Average Treatment Effect. Most of the times Stratified Propensity Score Analysis is used on non-randomized experiments where treated and non treated units have either observable or hidden relational nature; this is quite always the case of epidemiological studies [5] and Social Network analysis [19], where units are strongly related to one another and this connections can play a role in the likelihood of them being exposed to a treatment.

Socials Networks have this important aspect: users are interconnected under concepts like friendship, co-working, or (unidirectional) followship. Our extension of the concept of treatment and effect to social media, inspired by the work of Olteanu et al. [19] takes the form of *personal user experiences* as treatment and their outcomes as effects; we define *outcomes* as the words or expressions treated users are more likely to use in the future depending on their exposure to a particular life situation.

It goes without saying that to pursue this search for relations among experiences and outcomes we must turn to the Social Media world. Every day and hour—and sometimes minute—people disclose information to the free world through Social Networks. People use public posts to share what they are doing, eating, thinking; they sometime narrate past experiences, state or ask for medical condition, business and law situations, family and relationships events. Today's most massive Social Networks have made their way into people's lives—and smartphones—starting from few simple possible action, the main being posting broadcast *messages* under the most various names, commonly referred to as *status* or *post*, and *tweet* for Twitter, which have the particularity of being of no more than 140 characters and more than any other social, sets up an environment of short, concise statements, enhanced by *replying* or *retweeting* to other's posts creating a tree of asynchronous conversations [3].

2.1 Relational Propensity Score Estimation

What we want to address is a fair question: to which extent does an *individual-level causation* analysis discovers the potential outcomes of someone's personal experience, if we ignore network influences? How significant are our relations and ties in social networks for our choices after we find ourselves in a situation we might share with others?

We are then proposing a new approach to Propensity Score estimation that makes use of a Relational Classifier and then compare the resulting stratification of the population with one learned by a common classifier. As for the case being, as we will describe later, many times the observed covariates $\mathbf{x}$ may not be enough to explain the stratification in propensity of the population.

3 RELATED WORK

Propensity Score has been initially introduced and largely discussed by Rosenbaum and Rubin [20, 21]; it has been extensively used in all sorts of observational data analysis, and the literature provides a massive quantity of work on its use [12] for matching [6, 8], stratification and estimation of the Average Treatment Effect in observational studies [13, 17, 18], mostly on epidemiological studies [5]. Although treatment regimes need not be binary in practice, the propensity score methods are generally confined to binary treatment scenarios. Two possible exceptions have been suggested for ordinal and categorical treatments [16]. Propensity Score plays a fundamental role in assessing the generalizability of results from the observed population, or a randomized experiment, to a target one [23].

The motivation for our work comes mostly from Olteanu et al. [19] in *"Distilling the Outcomes of Personal Experiences: A Propensity-scored Analysis of Social Media"*, that initiated the problem of discovering experience outcomes from Twitter data. The work that they conducted mines potential outcomes from common users experiences, in trying to find the words people are more likely to use, or situations they are more likely to refer to, when they have been exposed in the past to some common scenario (e.g. is a user more likely to talk about *burning fat* if she has in the past stated its intention or need to *loose belly fat*?). This leads to a large variety of outcomes that we will borrow, together with the experiences, to set our work's baseline in Section 6.1.

Previous effort to study the effects of network relations has been conducted by Arbour et al. [1] in *"Inferring Network Effects from Observational Data"*, that contributed with the development of a tool called *Relational Covariate Adjustment (RCA)*, a general method for estimating causal effects in relational data.

4 RELATIONAL DATA STRATIFICATION

We introduce here the concept of Relational Data Stratification in which we propose a new approach to the Stratified Propensity Score Analysis altering the learning process for the Propensity Score estimation substituting the classical classifier with a relational one. A Relational Classifier infers the classification for a node which is part of a network from its neighbors. We use a *Weighted-vote Relational Neighbor Classifier* trained with 10-fold Stratified Cross Validation to preserve the user distribution between treatment and control group across folds [19].

Learning the Propensity Score estimation from a Relational Classifier means that we observe the Average Treatment Effect on *strata* of the population that have *similar propensity of being exposed to* *the treatment.* This propensity is not just inferred by the covariates **x** of each treatment unit, but from the influence of the network it belongs to and specifically its neighbors. This aspect can be crucial especially in observed social data, when selection bias [15] and self-selection [14] can significantly influence a unit's propensity of exposure to a treatment. Under this assumptions, we can dare to rewrite Equation (1) as

$$e(\mathbf{x}) = \Pr(z = 1|\mathbf{x}_1, \ldots, \mathbf{x}_d) \tag{3}$$

where $\mathbf{x}_i$ are the neighbors, in the network, of the unit $\mathbf{x}$, and $d = \text{degree}(\mathbf{x})$ is the number of links connected to $\mathbf{x}$.[1]

Our experimental work will focus on Social Network data to detect shifting in stratification from the Propensity Score estimation; to the best of our knowledge, this is the first attempt in making use of a Relational Classifier for Propensity Score Estimation.

5 DATA DESCRIPTION

Our first approach was to replicate the work in Olteanu et al. [19] to construct a baseline; their dataset[2] consists in a collection of tweets categorized by target experiences. The goal of the work was to mine users personal experiences—the treatment—as real life situations, medical and family condition, social issues; then derive frequent outcomes from those experiences through a stratified Propensity Score analysis. The set of experiences has been drew analyzing major search queries categorized by semantic domains (Olteanu et al. [19, Section 3]). The available tweets correspond to the set of *all tweets* that each treated user posted in English from March 1 to May 31, 2014. Treated users are identified by matching a set of words (e.g. *"belly fat"*) on all public tweets in April 2014. The original work resulted in about 20,000 potential outcomes based on concepts repetition among the tweets. Given the huge amount of data present in the released dataset, we extracted a feasible amount of tweets belonging to the same category *"Society Issues"*. We collected over 100,000 tweets for this category, resulting in the timelines of 320 treated users. Our collection of tweets is a subset of the original one—about 65% of the released tweets—due to the unavailability of some tweets more than three years later. We also took the list of the most 10 significant outcomes for our category.

We replicated what done in this previous work in the attempt to reform the Propensity Score estimation and observe variations in the population stratification. For our bad we could not proceed in the task as the network we reconstructed among Twitter users was almost absent. For this reason we move to a different dataset of Twitter data that would allow us to observe some sort of interconnection among the users.

We continued our work on a collection of political tweets surrounding the 2016 U.S. Presidential Elections. The dataset spans temporally from August to November and counts more than 24.6 million tweets from 4.1 million distinct users with an elevated percentage of retweets of highly followed and politically–oriented profiles' tweets. This drove us in building a *network of retweets* where a user x has an incoming edge from user y if x has ever retweeted a y's tweet; this convention might seem unusual, but is meant to represent a possible influence in x's propensity of being exposed to politics discussions when he/she retweets something political from y. This representation of the network will do so that our Relational Classifier will infer x's propensity from its neighbors.

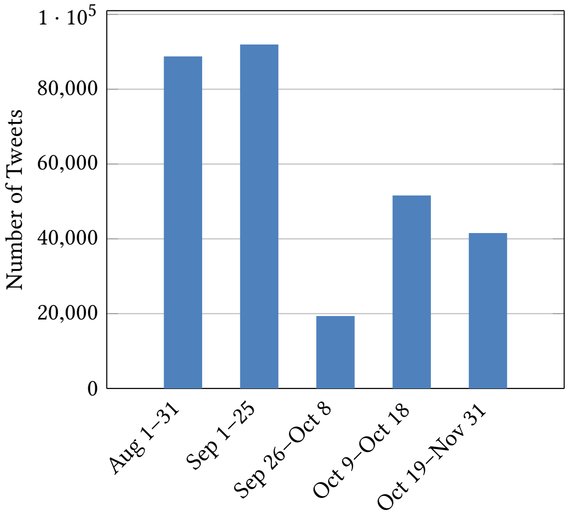

Figure 1: Tweets distribution over time, split according to the three Presidential Debates.

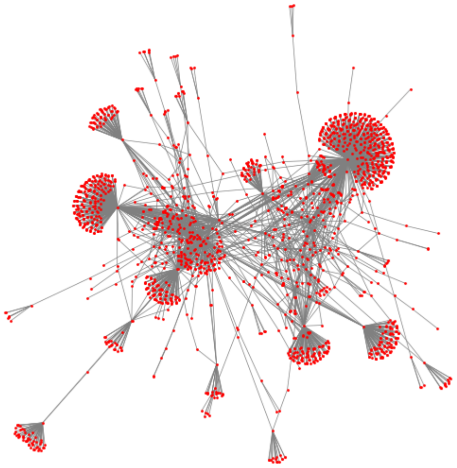

Figure 2: Snapshot of the retweet network among the users.

In the quest for a treatment–outcome relation in the dataset, we analyzed its distribution over time. We realized that a very diffused trending among tweets were talks about Presidential Debates; for this reason we looked for a possible treatment that could have influenced people into talking about debates. Figure 1 shows how tweets are spanned over time.

[1] Note that if the graph is directed, then we consider d as the in-degree.

[2] The set of tweet IDs from the Olteanu et al. [19] has been made publicly available by the authors at https://www.microsoft.com/en-us/download/details.aspx?id=54927

5.1 Data preprocessing

The first step is to preprocess data; as data comes from user–generated content and specifically text, this is a fundamental task to turn human written text into valuable information. For this task we make use of the Python Natural Language ToolKit library. The operations we perform are:

data cleaning removal of all records with missing or unexpected label

entities removal all the characters and strings that are not of interest; these include URLs, HTML tags, all punctuation and non–printable characters

data standardization it is necessary to bring data to a standard form, so that different variations of a word can be compared and matched; this process includes different steps:

- spell checking, to correct all misspelled words
- stemming, the process of reducing words to *stems* or *root form*, including verbs to present tense, plurals removal, root extraction from derived words
- stopwords removal, as they are not relevant to the analysis being
- shortening of all elongated words—like *looooong*—which are common in Social Network data as people use them to show feelings or give relevance to something [4]

5.2 Building the network

To construct a network among users we choose to link users based in their reciprocal retweets as described before. Given the huge amount of data — more than 3.4M retweets — we needed to sample the network in a way that allows us to work with a reduced dataset still being representative of the original one.

To ensure that we still have enough connections among the nodes we started sampling highly connected nodes — which count for about 1% of the dataset — and than retrieving their neighbors among all the remaining users up to a distance of 3 hops. This introduces a *selection bias* in the way users are exposed to the treatment [23] in addition to the *self–selection bias* which is phenomenon that can not be ignored in this type of data [14].

This sampling approach led to a particular network structure shown in Figure 2.

6 EXPERIMENTAL SETUP AND RESULTS

We introduce briefly how we implemented our baseline to pose the foundations of the framework that we will use.

6.1 Baseline implementation

To build this baseline we proceeded with steps borrowed by the framework described in Olteanu et al. [19, Section 4] looking forward to produce similar results that we shall later be able to compare.

(1) We first build user timelines grouping the tweets per user and sorting them according to the creation timestamp.
(2) We then process each user timeline to create two different timelines per user: one recording all the *first* and *last occurrences* for each outcome.[3]
(3) We identify *treatment* and *control* groups as the users that manifested or not a chosen target experience. Specifically, all users mentioning the experience *"losing belly fat"* formed the treatment group, while all the other users in the same category formed the control group.
(4) We constructed an array representation of the form:
$$H_{T_f} = [e_1, \ldots, e_n] \tag{4}$$
where e_i is 1 if the first occurrence of event i precedes the target experience—or an arbitrary event for the users in the control group that do not have the experience—and 0 otherwise. The same has been done for the timeline of the last occurrences.
(5) We learn the Propensity Score estimator for the *first occurrences* timelines using a Perceptron algorithm with a stratified 10–fold Cross Validation, ensuring that in each strata there is the same distribution of users in the treatment and control group.
(6) With the results of the Propensity Score estimation, we stratify users into 10 quantile–based strata in which users have similar propensity of reporting a predicted outcome.

Aside from the skewing produced by missing data from the original dataset, we can say that the values are still comparable to the ones reported by Olteanu et al. [19, Table 3]. From the non–relational Propensity-score stratification we obtain one single stratum containing almost all the population; this might be due to the fact that the Propensity Score estimation with Equation (1) including only the observed covariates $\mathbf{x}$ without users relations ignores the possible influence of a node's propensity to be exposed to the treatment from the network.

6.2 Working with Election Data

Since we were not able to expand the work from the baseline with a Relational Classifier due to the lack of connectivity among the users, we replicate the same framework on the Election data. Observing the data distribution shown in Figure 1, we selected time intervals according to the three Presidential Debates—September 26, October 9 and October 19, 2016. We then further divided the time frame to group data into three different periods:

(1) the debates period, from September 26 up to Election Day and further in later November; this is the period in which we will look for treatment effects.
(2) the month of September, pre–debates, in which we want to identify a possible treatment.
(3) a *control* period, the month of August, that we will use to learn the Propensity Score estimator, using pre–occurrences of the effects.

In mining possible treatment and outcomes, after the data pre–processing step, we perform a frequency analysis of expressions in tweets from early September (for the treatment) and from the first Presidential Debate on September 26 to Election Day for the

[3] The *outcomes* are tokens in the form of *unigrams* or *bigrams* like *"burn"* or *"work-out_lose"*.

Stratum	PS interval	Stratum Population	Effect %
1	0.26 — 0.30	5023	59.18
2	0.30 — 0.35	0	0
3	0.35 — 0.40	0	0
4	0.40 — 0.45	0	0
5	0.45 — 0.49	0	0
6	0.49 — 0.54	0	0
7	0.54 — 0.59	0	0
8	0.59 — 0.64	0	0
9	0.64 — 0.68	0	0
10	0.68 — 0.73	171	56.82

Table 1: Population stratification in the 10 quantiles of the Propensity Score learned with a binary classifier.

Stratum	PS interval	Stratum Population	Effect %
1	0.00 — 0.10	1193	35.04
2	0.10 — 0.20	8	12.50
3	0.20 — 0.30	46	30.43
4	0.30 — 0.40	34	55.88
5	0.40 — 0.50	62	56.45
6	0.50 — 0.60	3092	25.13
7	0.60 — 0.70	119	68.07
8	0.70 — 0.80	77	83.12
9	0.80 — 0.90	62	75.81
10	0.90 — 1.00	501	42.32

Table 2: Population stratification in the 10 quantiles of the Propensity Score learned with a Relational Classifier.

outcomes. After a human judgement on the obtained pattern, we discovered a potential treatment–effect pair. We assume that there could be a *"causality"* in talking about Presidential Debates in subjects that mentioned being interested in the campaign in early September.

Given the assumption above, we consider as exposure to the treatment tweeting about the upcoming campaign in the *treatment period*, and as effect the repetitive posting about political topics from the debates. Following what implement for the baseline we recollect data from the dataset and reconstruct the timeline for each user in a way similar to Equation (4):

$$[user, e_{\text{prev}}, t, e_{\text{after}}] \tag{5}$$

where e_{prev} and e_{after} are binary values indicating the presence of the effect respectively before and after the user shows exposure to the treatment t.

We use the representation in Equation (5) to learn the standard Propensity Score estimator as described in Section 6.1. The non–relational Propensity Score stratification is shown in Table 1. We can observe that all users are concentrated in the lower and higher quantile. This result is no particularly different in distribution from the one obtained in the baseline except for the highly separated upper and lower strata that can be caused by the fact that we are now considering only a binary outcome.

6.3 Relational Propensity Score

We now need a method to learn a Propensity Score estimator in a way that we can consider the network structure as a predictor. We adopted a *Weighted–vote Relational NeighborClassifier* trained with a stratified 10–fold Cross Validation and extract the propensity score value for each user. We used the classifier provided by NetKit, a Java based toolkit offering several useful functions to work on relational data. We expect these values to have an higher variance with respect to the non–relational ones and a significant shift in the population stratification given to the fact that the scores are inferred from an higher dimensional predictor.

The new stratification is reported in Table 2. With respect to the results described in Table 1 and obtained with the non relation classifier, the data are distributed among all the strata. With a less unbalanced stratification we can better observe differences in how the Average Treatment Effect varies from subpopulation with different propensity.

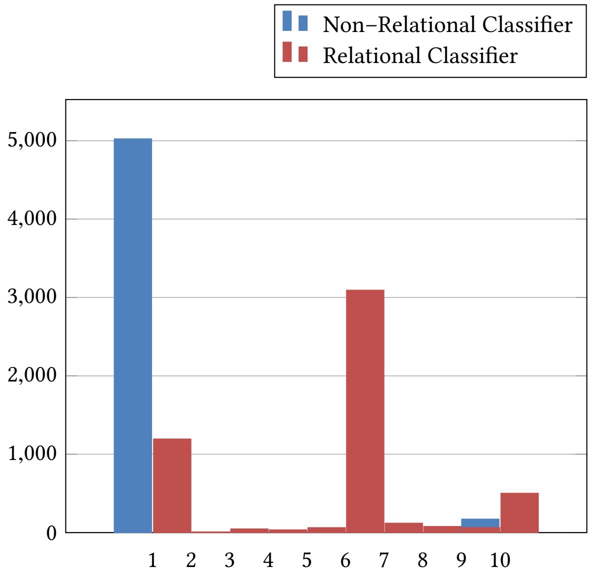

Figure 3: Comparison between the different distribution of the population among different strata with Non–Relational (blue) and Relational (red) classifier for Propensity Score estimation.

7 DISCUSSIONS AND FUTURE WORK

The Stratified Propensity Score analysis has been proven to balance the distribution of treatment and control groups in strata of units that are homogeneous in their propensity to be exposed to the treatment. Propensity Score based stratification allows not only to reduce selection bias and confoundness in observational data, but to assess the generalizability of the results from randomized trials to a target population. In real world data treatment and control group can have large differences in their observed covariates leading to a biased evaluation of the treatment effect.

In observational, non-randomized studies, the differences in observed covariates can produce uncontrolled bias in a unit membership to either the treatment or the control group; The point in observational studies is that, unlike an experiment, there is no systematic method to maintain a control group such that it is not biased the way the treatment is assigned to units, that can be the result of self-selection or systematic judgement by the researcher. This motivates our intent in including network information in propensity score estimation on relational data learning the estimation with a relational classifier.

We worked on Twitter observational data considering treated users that showed interest in the 2016 Presidential campaign with respect to their later behavior in talking and retweeting about the Presidential Debates. We have learned a Propensity Score estimator both with the only covariates of the units as in the classic approach, and with a Relational Classifier. The Relational Propensity Score analysis led to significantly different results in the stratification of the population and a more fine–grained evaluation of the Average Treatment Effect.

Having assessed the relevance of introducing network influence in the Propensity Score estimation for relational data we can motivate future research work towards the assessment on how the Treatment Effect evaluation can benefit from a differently stratified population depending on the relational nature of the data.

8 ACKNOWLEDGEMENT

A great thanks goes to the BigData Laboratory[4] of the Politecnico di Torino, in the person of the Prof. Paolo Garza, that allowed us to use their powerful resources that was fundamental to work on huge data.

REFERENCES

[1] David Arbour, Dan Garant, and David Jensen. 2016. Inferring network effects from observational data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 715–724.

[2] Peter C Austin, Paul Grootendorst, and Geoffrey M Anderson. 2007. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in medicine* 26, 4 (2007), 734–753.

[3] Danah Boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*. IEEE, 1–10.

[4] Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllll-lll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 562–570.

[5] M. Alan Brookhart, Sebastian Schneeweiss, Kenneth J. Rothman, Robert J. Glynn, Jerry Avorn, and Til Strmer. 2006. Variable Selection for Propensity Score Models. *American Journal of Epidemiology* 163, 12 (2006), 1149–1156. DOI: http://dx.doi.org/10.1093/aje/kwj149

[6] Marco Caliendo and Sabine Kopeinig. 2008. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys* 22, 1 (2008), 31–72.

[7] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2098–2110.

[8] Rajeev H Dehejia and Sadek Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *The review of economics and statistics* 84, 1 (2002), 151–161.

[9] Ralph B dfiAgostino. 1998. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 17, 19 (1998), 2265–2281.

[10] Dean Eckles, René F Kizilcec, and Eytan Bakshy. 2016. Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7316–7322.

[11] Sharad Goel, Andrei Broder, Evgeniy Gabrilovich, and Bo Pang. 2010. Anatomy of the long tail: ordinary people with extraordinary tastes. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 201–210.

[12] S. Guo and M.W. Fraser. 2014. *Propensity Score Analysis: Statistical Methods and Applications*. SAGE Publications. https://books.google.com/books?id=TT8sBQAAQBAJ

[13] Jinyong Hahn. 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* (1998), 315–331.

[14] Glenn W. Harrison, Morten I. Lau, and E. Elisabet Rutstrm. 2009. Risk attitudes, randomization to treatment, and self-selection into experiments. *Journal of Economic Behavior & Organization* 70, 3 (2009), 498 – 507. DOI: http://dx.doi.org/https://doi.org/10.1016/j.jebo.2008.02.011 Field Experiments in Economics.

[15] James Heckman. 1990. Varieties of Selection Bias. *The American Economic Review* 80, 2 (1990), 313–318. http://www.jstor.org/stable/2006591

[16] Kosuke Imai and David A Van Dyk. 2004. Causal inference with general treatment regimes: Generalizing the propensity score. *J. Amer. Statist. Assoc.* 99, 467 (2004), 854–866.

[17] GW Imbens. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87, 3 (2000), 706–710.

[18] Jared K Lunceford and Marie Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine* 23, 19 (2004), 2937–2960.

[19] Alexandra Olteanu, Onur Varol, and Emre Kiciman. 2017. Distilling the Outcomes of Personal Experiences: A Propensity-scored Analysis of Social Media.. In *CSCW*. 370–386.

[20] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[21] Paul R Rosenbaum and Donald B Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association* 79, 387 (1984), 516–524.

[22] Donald B Rubin. 1997. Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine* 127, 8_Part_2 (1997), 757–763.

[23] Elizabeth A Stuart, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf. 2011. The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 174, 2 (2011), 369–386.

[4] https://bigdata.polito.it/