CSC2552 Assignment 1

Instructions

• Your work must be your own.
• This assignment is due at 9pm on Monday, February 17 (hard deadline).
• Your assignment must be typed. You may use any word-processing software you like (e.g. LaTeX).
• You may use whatever programming language and plotting libraries you like.

Problem 1

In our first discussion week, we read the widely-circulated paper “Fake news on Twitter during the 2016 U.S. presidential election” by Grinberg, Joseph, Friedland, Swire-Thompson, and Lazar. During our in-class discussion, some questions came up that weren’t directly addressed in the paper or the supplemental materials. Thanks to the era of open and reproducible science we’re living in, we can answer these questions ourselves. Working through this problem will give you a chance to handle some real-world data, reproduce a state-of-the-art result, and perform a follow-up analysis of it.

(a) Download the data from [https://doi.org/10.5281/zenodo.2483311](https://doi.org/10.5281/zenodo.2483311) (this is listed as reference 26 in the paper, and is mentioned in the Data and materials availability section at the end of the paper — this is often where pointers to replication materials will be.). Under the “Files” section on this page, there is a link that says “Download”. Unzip the archive on your hard drive. The archive has a lot of good stuff in it — browse around and look at all of the different datasets the authors assembled to do their analyses.

(b) Recreate Figure 1A of Grinberg et al. To do this, you will need the data in `figure_1/data/daily_counts_exposures.csv`. The x-axis information is in the `date` column, and the y-axis information, the percentage of exposures per news source type, is in the `pct_red`, `pct_orange`, `pct_black` columns. The authors visualized the data with a stacked histogram, but you can also draw a line graph if you find it easier (one line each for the black, orange, and red sources). Notice that the authors have provided replication code for the Figure in `figure_1/gen_fig1_plus_supporting.R`, so you can run this code for this question if you are familiar with R. If you aren’t, you can use whatever software you are comfortable with. Don’t worry about the “2016 election day” annotation or typesetting the dates nicely.

(c) In class, we discussed how the authors’ decision to report all of the results in percentages was usually fine, but it also obscured a few details. For example, our fearless presenters pointed out that there is a strange dip in Figure 1A on September 26. Some digging revealed that this was the date of the first Presidential debate. Is the dip on September 26 due to a decrease in fake news that day or an increase in normal news that day?

Let’s find out. Using the same data, create a version of Figure 1A where the y-axis is expressed in absolute terms, not relative terms. That is, the y-axis should be the raw number of fake news exposures per source that day. The x-axis is the same as before (the `date` column) and the y-axis information, the raw number of exposures per news source type, is in the `red`, `orange`, `black` columns. Again, you may use a stacked histogram or a line graph (with one line per fake news type) to visualize the data.

(d) Based on your plot in part (c), was the dip on September 26 due to a decrease in fake news that day or an increase in normal news that day? Explain in one sentence.
(e) What other features of the data stand out when you view the data on an absolute scale, as opposed to a relative scale? Explain in 2-3 sentences.

**Summary.** For this question, hand in a plot for part (b), a plot for part (c), and concise explanations to parts (d) and (e). At the end of the assignment, append the code you used to produce your plots.

**Problem 2**

In this question, we’ll explore ways of approximating experiments with big data through natural experiments.

(a) Explain the difference between a natural experiment and a randomized experiment in 1–2 sentences.

(b) As you’ve learned, natural experiments are a way to exploit random variation “in the wild” — but in order to do so, you must find a source of random variation. One common such source is the weather (i.e. sometimes it rains and sometimes it doesn’t, and this variation is usually external, or “exogenous” as economists say, to the social system being studied). Come up with a research question that you could approach by setting up a natural experiment where the weather is used as the source of as-if random variation, and briefly describe the natural experiment setup you have in mind.

(c) Briefly describe a “counting things” approach to the research question you posed in the previous question.

(d) How do these two approaches compare? Would you prefer the natural experiment or the counting things approach? Explain in 2–3 sentences.

**Problem 3**

In Section 2.3 of *Bit by Bit*, Salganik enumerates ten common characteristics of big data, some generally good and others generally bad. Come up with one additional positive characteristic and one additional negative characteristic of using big data for social research. For each one, explain the characteristic and why it is generally good/bad in 1–2 sentences.

**Problem 4**

Enumerate Salganik’s ten common characteristics of big data, and for each characteristic mention one of the four papers we read in the Observational Studies section of the course that it applies to. Explain why it was important to the paper you chose in 1 sentence. Refer to Papers 1–4 as “Grinberg et al.”, “Obermeyer et al.”, “Bakshy et al.”, and “Kleinberg et al.”, respectively.