# Schenk_Project_Proposal_9.30

October 1, 2019

Marie Schenk

First Draft: 9/24/2019 This Draft: 9/30/2019

GitHub repository: https://github.com/mhschenk/1030CCES

## 1 Describe the problem you want to solve.

Who uses social media to post about politics? Is there significant geographic variation in rates of posting about politics? Are people more likely to post about politics in the context of highly competitive political races, or are general demographic characteristics like age and education more predictive?

The target variable is actually four variables, which I intend to collapse into an index. They are related questions about activity on social media related to political content. They ask, have you (1) posted a story, photo, video, or link about politics; (2) posted a comment about politics; (3) read a story or watched a video about politics; (4) followed a political event; (5) forwarded a story, photo, video, or link about politics to friends? I will have to perform some additional exploratory analysis before deciding exactly how to collapse these. Unsurprisingly, many more people have read a political story than have posted their own political content, so it may be necessary to separate that subquestion.

This could be a regression or a classification problem, depending on how I recode the target variable. If I make it a dummy variable (respondent did any of those activities), it will be a classification problem. If I make it a scale (how many of these activities respondent did), it will be a regression problem. How I code it will depend on how many people have done just one of the activities. If most people have done all of the activities, I will make it a dummy variable, because the variation in the scale will not be very meaningful.

This project relates to my dissertation in political science, which is about how online spaces convened for non-political purposes become politicized. I am looking at this in the context of what Jane Mansbridge calls "everyday political talk." Everyday talk may not be overtly political, in that it does not name public policies or political institutions, but nevertheless address issues that impact public life. Everyday talk and public deliberation have been a topic of interest for political scientists for years, but most work focuses on spaces such as cafes or public meetings, not virtual spaces. The advent of social media has dramatically changed the structure of public life; many of the activities that used to occur in public spaces, such as coffee shops, now take place in virtual spaces, like Twitter threads. This project contributes to the larger project of my dissertation by helping me understand if online discussions are similar to other, more widely studied forms of political engagement. It is well established in political science, for example, that people who vote tend to be older and more highly educated than average. We also know that voting rates go up when highly visible, highly competitive elections are taking place, although the mechanism for that is somewhat unclear. In what ways is talking about politics online similar to voting, and in what ways is it different? This will help me create expectations for my future work, based on what existing theories I think are most likely to apply to online political discussions. Since online discussion is an understudied aspect of political life, understanding how it compares to better understood phenomena is a valuable starting point.

I am particularly interested in exploring the differences between a data-science approach to answering this question and a political science approach. As a conclusion to this paper, I would like to reflect upon methodology, in addition to my findings. Since I have subject-area knowledge about politics, I would like to bring it to bear by framing the discussion in this way.

## 2 Describe the dataset.

The Cooperative Congressional Election Study, or CCES, has been conducted annually since 2006. The survey has a unique design. One portion, called the Common Content, is asked of every person who takes the survey. The Common Content consists of questions that are of broad interest to social scientists studying political phenomena, including demographic information, respondent's party identification, and political knowledge. It also includes verified voter information. Using a commercial service called Catalist, the CCES matches respondents to public voter file data. This is especially important because voting is a socially desirable activity, and people may lie on a survey and say they voted, because they know it is a good thing, when in fact they did not. In addition to the Common Content, teams of researchers across the country can put their own modules on the survey. These questions are asked to a subset of 1,000 survey respondents. This design gives many researchers the opportunity to ask their questions to a sufficiently large survey to produce a nationally representative sample. The much larger number of respondents who receive the Common Content questions allows for a sample that is not only nationally representative, but also representative of smaller geographic units, allowing more precise study of Congressional races.

The dataset I am using is the Common Content from the 2018 round of the CCES. There are 524 variables and 60,000 observations in this data file. The 2018 Common Content is well described in the guide available for download on the Harvard Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910/DVN/ZSBZ7K.

The CCES is a commonly used datasource in political science, espeically for work that requires a representative sample of a geographic area smaller than the whole country. Alexander Hertel-Fernandez, Matto Mildenberger, and Leah C. Stokes (2019) use questions about opinions on policy issues from the 2016 CCES to assess public opinion in various legislative districts. They compare these results to a second, specialized survey they conducted of legislative staffers, to see if the staffers had an accurate understanding of the opinions and needs of the consituents in their district. Danny Hayes and Jennifer Lawless (2015) merge 2010 CCES data on respondents' political knowledge with their own data on the availability of local newspapers. They use several items from the CCES, including questions that ask the respondent to rank the quality of their House candidate, describe the ideological position of the Democratic and Republican candidates running for House in their district, and whether the respondent plans to vote in the upcoming House election. They find that democratic engagement, as measured by these items, declines in districts with uncompetitive races and without robust local news coverage.

There are many other political science publications that make use of the CCES. These are particularly good examples because they exploit the facet of the survey that is most unique–the ability to study public opinion in individual Congressional districts. I also plan to use the survey to its greatest advantage by looking at geographic variation in the way people engage with politics online.

## 3 Preprocess the data

See code in 'src' folder in GitHub repository.

[ ]: