Using Longitudinal Social Media Analysis to Understand the Effects of Early College Alcohol Use

Emre Kıcıman[†] and Scott Counts[†] and Melissa Gasser[‡]
[†]Microsoft Research, [‡]University of Washington
emrek@microsoft.com, counts@microsoft.com, mlgasser@uw.edu

Abstract

While college completion is predictive of individual career happiness and economic achievement, many factors, such as excessive alcohol usage, jeopardize college success. In this paper, we propose a method for analyzing large-scale, longitudinal social media timelines to provide fine-grained visibility into how the behaviors and trajectories of alcohol-mentioning students differ from their peers. Using propensity score stratification to reduce bias from confounding factors, we analyze the Twitter data of 63k college students over 5 years to study the effect of early alcohol usage on topics linked to college success. We find multi-year effects, including lower mentions of study habits, increased mentions of potentially risky behaviors, and decreases in mentions of positive emotions. We conclude with a discussion of social media data's role in the study of the risky behaviors of college students and other individual behaviors with long-term effects.

Introduction

College is an important transition period in the life of young adults (Arnett 2000; Lu 1994). Success in college can predict individual career success, career happiness, and economic achievement (Griliches and Mason 1972), and is of broader societal importance as well, as high rates of post-secondary degree holders are believed to drive income levels and other measures of macro-economic growth (Krueger and Lindahl 2000). Yet, approximately one in three college students leaves without earning a degree (Shapiro et al. 2015). Many factors, including adjustment challenges, family responsibilities, financial pressures, and individual behaviors can jeopardize college success (Pantages and Creedon 1978; DeBerard, Spielmans, and Julka 2004; Tinto 1987).

One factor negatively associated with college success is excessive alcohol consumption. On average, college students drink more than their same age non-college peers (Schulenberg et al. 2001). This is a persistent public health issue, including high rates of binge drinking and wide-ranging consequences such as hangovers, lowered academic performance, DUI arrests, risky sexual behavior, sexual and other assaults, and alcohol-related injuries and deaths (Johnston et al. 2011; Nelson et al. 2009; Hingson, Zha, and Weitzman 2009; Hingson 2010; Wechsler et al. 1995; Musgrave-Marquart, Bromley, and Dalley 1997).

Given this wide range of effects of college drinking, research methods to address the issue will benefit from incorporating as many aspects of the lives of college students as possible. This paper utilizes a large-scale, longitudinal social media data set to provide a view into a broader group of college students which can supplement large sample longitudinal self-report research, such as the Monitoring the Future study (Johnston et al. 2011). Given the high frequency of posting, we find that social media streams provide a granular and *in situ* reporting of students' topical interests, behavioral patterns and activities, including risky behaviors such as alcohol use, and the contexts and consequences of these activities which are difficult to capture at scale through existing methods.

We thus generated a multiyear dataset of the social media timelines of 63,387 students entering college in 2010, capturing their timelines from August 2010 through May 2015. We focus on the effect of drinking early in college, using propensity score analyses to identify increased likelihood of mentions of scholastic and social outcomes, such as study habits, social relationships, and even criminal activity of those who mention drinking during their first semester versus those that do not.

The primary contribution of this paper is a longitudinal analysis of long term effects of drinking early in college that utilizes propensity scoring stratification to minimize confounding factors. Multiyear outcomes are associated with early alcohol exposure, including fewer mentions of study habit and social relationships, and sustained increases in mentions of alcohol and other risky behaviors including criminal activity. This analysis fills a gap in the literature on alcohol use in college by showing that social media-based measures of alcohol use early in college can be used to assess the future likelihood of negative outcomes later in college.

Background and Related Work

Challenges in the College Period

College is the first time that many young adults live outside the home and a time when "emerging adults" are developing new social, cognitive, and awareness skills (Arnett 2000). While many college students successfully manage this transition, graduating and emerging into adulthood, there are also many students who experience significant challenges. Studies going back decades find stubbornly high attrition rates, reporting that up to 40% of students leave without a

degree (Porter 1989; Pantages and Creedon 1978).

Many factors are associated with or predict academic performance and college attrition, including academic and social integration variables (Terenzini and Pascarella 1978; McKenzie and Schweitzer 2001). However, while prior academic experience and success are significant predictors, they do not account for the most success variance (McKenzie and Schweitzer 2001; Wolfe and Johnson 1995). Instead, many studies have found that psychological measures are strongly associated with college success. Lecompte *et al.* find that expectation of academic success has a positive association with actual academic success and low attrition (Lecompte, Kaufman, and Rousseeuw 1983). Gerdes and Mallinckrodt find that emotional and social adjustment better predicts attrition than prior academics (Gerdes and Mallinckrodt 1994).

Studying factors related to academic success and challenges in college can help drive interventions to better support students in need, including early intervention programs in high schools, appropriate counseling services, financial aid, academic services, and programs that aid student socialization and community integration (Clark and Halpern 1993; Valentine et al. 2009). Methodologically these studies use surveys primarily, which are often limited to single institutions, rely on participant recall, and are subject to response biases. Ideally, social media-based research can complement existing methods, deepen understanding of real experiences given its in situ nature, and provide a timeliness that can be leveraged for intervention.

Research methods to study college student behaviors, stressors, and related outcomes, have to date been largely limited to self-report surveys and interviews, and in-person or deliberate recruiting research. While such methods have many benefits, including the ability to gather rich, open-ended detail, they also have significant limitations, including having relatively small sample sizes from single-institution studies, limiting generalizability of findings.

College Students and Social Media

Approximately 80% of young adult internet users (ages 18-29) use social networking sites (Facebook, Twitter, Pinterest, Instagram, Tumblr), with 27% on Twitter specifically (Duggan and Brenner 2013). As usage of social networking sites among college students has grown, content posted by students has become a significant information source, albeit biased, about students' fine-grained activities and interactions.

Many studies of social networking site usage among college students have found that social networks play an important role for students in this transition period. Social media provides informational and social support, a source of connection to other students and university community, as well as connection to family and friends (Ellison et al. 2011; Ellison, Steinfield, and Lampe 2007). Gray, Vitak, Easton and Ellison study social adjustment to college in freshman year using online social network measures such as number of friends, and find positive relationships between engagement and collaboration with other students through Facebook and measures of social support and adjustment (Gray et al. 2013).

Previous research has shown that measures derived from social media posts can predict alcohol related behavior. Marczinski et al. (Marczinski et al. 2016) assessed alcohol-related Facebook activity, finding it predictive of measures of alcohol use quantity and frequency and risk of alcohol use disorder.Social media posts by others in ones' network has also been shown to predict later alcohol usage, particularly for male students. For example, undergraduate first years' exposures to alcohol-related content in their first 6 weeks predicted alcohol use six months later (Boyle et al. 2016). Building on this research, we examine downstream effects of alcohol use, measured through social media post activity.

Longitudinal Studies of Online Data

Longitudinal studies of online data, including social media data and search query logs, have proven effective in helping understand the behaviors of people in various situations. These studies have been targeted to explore and understand how situations evolve over time, identify predictive factors involved in positive and negative outcomes, and help identify at-risk individuals. For example, using search query logs, Paul et al. (Paul, White, and Horvitz 2015) characterize the information seeking behavior during various phases of prostate cancer. Fourney et al. (Fourney, White, and Horvitz 2015) align search query logs with the natural clock of gestational physiology of pregnant women to characterize their changing information needs. Althoff et al. study 5 years of fitness tracking data to better understand social influence on physical activity (Althoff, Jindal, and Leskovec 2017).

By mining social media, De Choudhury et al. (De Choudhury et al. 2013) find behavioral cues useful to predict the risk of depression before onset. Similarly, by leveraging these naturalistic data, prior work examined how dietary habits vary across locations (Abbar, Mejova, and Weber 2015); the links between diseases, drugs, and side-effects (Myslín et al. 2013; Paul and Dredze 2011); links between actions and outcomes (Kıcıman and Richardson 2015); and shifts in suicidal ideation (De Choudhury et al. 2016) . Olteanu et al. demonstrate propensity scored analysis of social media timelines to understand outcomes across a broad set of domains (Olteanu, Varol, and Kıcıman 2017).

Methods

Identifying Students Entering College

To identify students entering college, we first use high-recall, low-precision phrase matching to select a set of tweets indicating the author may be starting college. Then, we apply a high-precision text classifier to identify individuals who are very likely to be starting college. Thus, after retrieving a large set of tweets related to college attendance, our high-precision classifier distinguishes between tweets indicating the author is starting college ("Can't wait to start college next week") and tweets that do not ("Can't wait for college football").

We create our high-recall, low-precision phrase set using an iterative keyword generation procedure of tweet retrieval, evaluation, and keyword expansion. We identify a list of 87 phrases related to college attendance, and extract 639k tweets that match these keywords during a 5-month period in the fall of 2010 (August-December 2010) in our organization's archive of the Twitter firehose, restricting our analysis to only

Table 1: Top college-attendance keywords and paraphrased examples of tweets passing or failing our high-precision classifier.

Keyword Phrase	Positive example	Negative example
day of college	First *day of college* is tuesday. i am not ready yet :(	my little sis is leavin for her first *day of college* @ UMich.. * thos arent tears *
college tomorrow	Last day on the beach! :(headed to *college tomorrow* :)	so bummed don't even want to go visit my dream *college tomorrow* might just stay in bed ...
start college	woah can't believe I *start college* tomorrow. today is the last official day of vacation.	when you *start college* are you going to focus on that
going to college	@user nervous *going to college* is gonna be big change for me :D	old guy told me if I plan on going to college, Ill need better posture. wth
my first college	Had *my first college* class!! its official i am a college student	at *my first college* I got in trouble because I did not go to church
my first semester	Just paid tuition for *my first semester* of university... my wallet is sad :(	rereading the essays when I did *my first semester* 3 years ago. my writing sucked LOL

Twitter users with English-language profiles. Table 1 lists the number of users and tweets identified by top phrases.

To build our classifier, we labeled a selection of 1000 tweets from our initial retrieval using Mechanical Turk judges. For each tweet, we asked three judges to indicate whether the author was in or soon to be in college; not in college; or cannot tell. For in-college tweets, we asked whether the author appeared to be in or entering their first year. We remove tweets where a majority of judges disagree on the label and downsample to create a balanced dataset of 634 tweets by first year students and others. Using unigrams, bigrams, and part-of-speech tags as features, we train a logistic regression classifier with 10-fold cross-validation and achieve a classification f-score of 0.78 (AUC=0.85). We set a high acceptance threshold to trade off recall for high precision. Our final classifier achieves a precision of 88% and a recall of 25%. Applying this to our dataset we identify 63,387 users likely to be starting their first year of college in 2010.

We retrieved from our Twitter firehose archive all tweets (including retweets) by these users during the almost 5-year period from August 2010 through May 2015. This time period is intended to cover the full 4 to 5 year college tenure of students. This dataset contains 658,905,460 tweets by the 63,387 users over almost 5 years.

We note here that for privacy and ethics considerations, all tweet content used is publicly available, all analyses were conducted anonymously, and all results aggregated. Further, example tweet texts used in this paper for illustration purposes have been paraphrased and subsequently checked via Twitter search to ensure tweet authors are not identifiable.

Characterizing Tweet Content

Identifying Alcohol Usage We identify potential signs of drinking alcohol using a curated list of keywords associated with alcohol. Our keywords are developed by one of the authors, a research expert on alcohol usage and risky behavior, based on commonly used phrases and slang for drinking in addition to the most popular brands of beer and liquor and retail store sales by a market research firm.

Keywords were further refined through multiple iterations to identify additional terms and phrases commonly used in conjunction with these brands and drinking indicators, and to disambiguate from false positives. Our final keyword list consists of 111 words and phrases about alcohol use. A selection can be found in Table 3. We apply these keywords to all original tweets (not retweets) to identify users mentioning alcohol consumption.

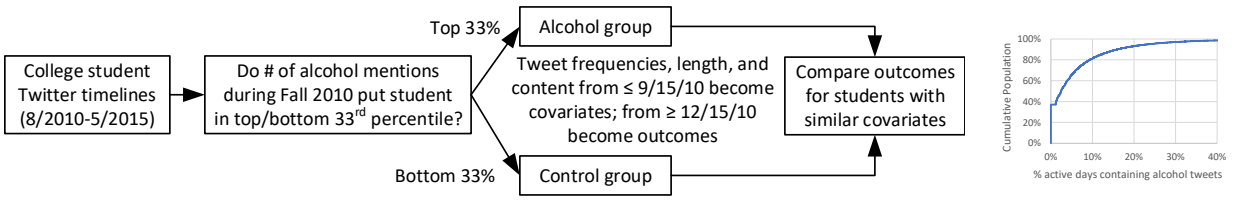
Identifying Topics Relevant to College Success To better understand the relationship between alcohol mentions and outcomes that might be affect college success, we seek to identify effects that are known to be linked to college success: peer group interactions, family responsibilities, study habits, negative academic outcomes, financial pressures and legal/criminal challenges. To bridge between these high-level concepts and the textual representation of social media timelines, we used Empath (Fast, Chen, and Bernstein 2016) to systematically generalize from a sample of seed words in our selected topics to related words in the same lexical category. Using Empath, we generate a short, non-overlapping list of 20-60 related words for each concept (Table 2). In addition, we also analyze the effects of alcohol mentions and Empath's built-in 195 human-validated topics, including additional risky behaviors, emotional topics, and other indicators of social and work-related topics.

Propensity Score Analysis

Our analysis goal is to understand the effects of experiences (e.g., drinking alcohol) mentioned early in a person's college career using observational study. Because we are interested in better understanding potential mechanisms and possible interventions, our goal is fundamentally one of causal inference. While we do not believe we can achieve the ideal identification of causal relationships, we can use methods borrowed from the causal inference literature to reduce the bias of naive correlational analyses. Prior research demonstrates the feasibility of this approach. For example, Eckles and Bakshy reduced bias in an observational study by 97% compared to a naive analysis, as measured against a gold-standard randomized field experiment, by conditioning on high-dimensional covariate data (Eckles and Bakshy 2017). Towards this end, we similarly condition on high-dimensional covariate data — the distribution of words used by individuals in the 6 weeks from Aug. 1 through Sep 15, 2010 — with the aim of reducing bias in our estimates of causal effects. Specifically, we apply a stratified propensity score analysis (Rosen-

Table 2: Topics linked to college success

Concept	Seed words	Final topic words	Example Tweets
Peer group interaction	friend, boyfriend, girlfriend	boyfriend, buddy, roommate, bandmate, fiance, *+20 more*	Yup buddy I found my bandmate!!
Family responsibilities	mother, father, brother, sister	mother, father, brother, stepdad, grandmother, *+21 more*	Thankful for my little bro and mom I have a sister #fact
Study habits	study, library, homework	study, library, math, tutor, textbooks, worksheets, *+49 more*	anyone that wants to study for history we're in the library but anyways ima off to study
Financial pressures	debt, student loans, loans	wages, afford, utilities, tuition, evicted, fees, *+28 more*	finally my wages wooo @anon its all about money. Im in debt. dont want more loans
Legal/criminal challenges	police, cops, jail, parole	cops, police, restraining, probation, rehab, *+15 more*	meeting my parole officer cops pulling out breathalyzer f***k we drunk



(a) Analysis procedure (b) Alcohol mentions CDF

Figure 1: (a) Schematic description of our analysis procedure.; (b) Population CDF of active days of alcohol mentions during Sep. 15-Dec. 15, 2010. 37% of users never mention alcohol during this period, while 19.7% mention alcohol on over 10% of the days they are active on Twitter.

baum and Rubin 1983), a method of conditional inference within the potential outcomes framework (Rubin 2011; Imbens and Rubin 2015) for causal inference.

In the potential outcomes framework, whether some experience "causes" an outcome, is computed by comparing two potential outcomes: one outcome $Y_i(T=1)$ after a person i has a target experience T [1], and another outcome $Y_i(T=0)$ when the same person in an identical context does not have the experience. The causal effect of T is then $Y_i(T=1) - Y_i(T=0)$. Of course, it is impossible to observe both $Y_i(T=1)$ and $Y_i(T=0)$ for the same individual i. In a sense, the problem of causal inference is a problem of missing data, and causal inference techniques attempt to address this challenge by estimating the missing counterfactual outcome for an individual based on the outcomes of other, similar individuals.

The stratified propensity score analysis estimates missing counterfactual outcomes by identifying matching subpopulations of individuals with similar distributions of covariates, but with differing treatment status. See Figure 1(a) for a high-level representation of the analysis steps. Conceptually, the idea is to find pairs (generalizing to groups) of individuals in the observational data whose covariates are statistically very similar to one another, but where one has received a treatment and the other has not. In this study, a person who is in the top 33% of the population, as measured by the number of alcohol-related tweets they post during the period Sep. 15-Dec. 15, 2010 is in the *alcohol group*, and a person who has not tweeted about alcohol is in the *control group*. Additional details are given Alcohol Mentions Section.

Matching of groups is achieved by estimating every individuals likelihood of being in the Alcohol group using a propensity score model. This is a learned function that infers likelihood of being in the Alcohol group as a function of a set of covariates (i.e., individual properties and past tweets that might influence both group status and outcomes). Individuals with similar propensity scores are grouped into strata. In aggregate, individuals within a strata are likely to have similar covariates, allowing us to isolate and estimate the effects of the treatment itself within each strata. Note that the primary purpose of the propensity score model is to identify groups of individuals with similar covariates—the accuracy of predicting group status is secondary. To ensure quality of counterfactual estimates, the method drops strata that have either too few Alcohol or too few Control users, and aggregates outcomes across remaining strata.

We reserve tweet histories from Aug. 1 through Sep. 15 2010 as covariates in our stratified propensity score analysis, and measure outcomes from December 15, 2010 through 2015. These covariates in our study consist of tweet frequencies, tweet lengths, and word distributions. We gather these covariates from the beginning of our data set (Aug. 1, 2010) until Sep. 14, 2010. To ensure quality covariate matching, we remove users below the 50th percentile of tweet volume and above the 99th percentile of tweet volume. We measure outcomes (e.g., word and topic distributions) beginning from Dec. 15, 2010, over a 28-day window, and slide this window over time to allow us to characterize dynamically

[1] In medical and social sciences literature, the target experience is often called the *treatment*, and is compared to a *control* or *placebo* experience. We will use the term Alcohol group and Control group in this paper

varying effects. Our word distributions (both for covariate and outcome analysis) are characterized as the empirical, unsmoothed word likelihood and include the top 50k unigrams in our corpus for covariates, and a fixed vocabulary of the top 10k unigrams for outcome words, not including URLs and user mentions. We do not remove stopwords, stem or normalize the text, and use whitespace and punctuation to identify word-breaks. We combine outcome word likelihoods for all words in a given topic to generate the total topic likelihood.

We implement our high-dimensional propensity score analysis as a logistic regression with 10-fold cross-validation. Our analysis divides users into 100 strata, removes strata with either or both too few Alcohol or too few Control users. In practice, this removes the lowest-propensity strata and the highest-propensity strata, leaving the middle strata in these analyses. The outcome differences in these remaining strata are weighted according to the Alcohol population distribution and combined to estimate the average treatment effect on the treated population (Alcohol group).

In terms of assessing statistical significance of differences between the Alcohol and Control conditions, we include in the relevant figure captions p-values and Cohen's d effect sizes. These were obtained by comparing monthly aggregates of topical word likelihoods over all people in each of the two conditions. Thus these statistics are based on an N of 55 months rather than an N of tens of thousands of people in the study. This is therefore a conservative estimation of differences between the Alcohol and Control groups.

Data and Analysis Limitations

While social media data is recognized as a rich data source, capturing a wide array of information about both on-line and off-line human behavior, experiences and dynamics, we recognize the many factors that bias social media data sets and their representation of on-line and off-line events. These factors include population biases (Mislove et al. 2011; Diaz et al. 2016), self-presentation and behavioral biases (Kıcıman 2012; Gong et al. 2016), potential algorithmic confounding (Hargittai et al. 2010) and biases introduced by differing affordances in the underlying social platform (Malik and Pfeffer 2016). With a longitudinal study such as this, survival bias (who continues to tweet over time) is also a potential biasing factor. Furthermore, though we augment our quantitative analyses with qualitative sampling of underlying messages, we do rely significantly on various machine-learned classifiers and mappings, such as the propensity score estimator, and machine learned topic mappings. Systematic errors in these algorithms may lead to unexpected biases.

Our propensity score analysis may be affected by unobserved confounding variables. In particular, the effects we identify should be interpreted as being linked to the social and physical circumstances both on-line and off-line that lead individuals to post alcohol mentions. As our analysis only stratifies on potential confounds measured prior to Fall 2010, experiences occurring concurrently with our counts of alcohol mentions during the Fall 2010 are effectively unobserved confounders, and their influence will be entangled in our results. Similarly, it is possible that the Control group population contains people who mentioned alcohol prior to our observation period but not during their first college semester. Finally, our analysis assumes treatment effects on an individual are independent of others' treatment status.

College Student Timelines

Our data consists of 658M tweets by 63k users, recorded from Aug. 1, 2010 to May, 2015, representing approximately 12 messages per day per active user. Each tweet includes message text, creation date, user id, and profile location.

User characteristics

We briefly characterize of the gender and geographic makeup of the individuals in our dataset. The gender distribution, inferred from users' first names cross-referenced with United States Social Security Service records of annual births, indicates that our user base consists of a plurality of users with names identifiable as female names (40.6%). The remainder of accounts are split among male (29.8%) and users with names not identified with a gender (29.5%).

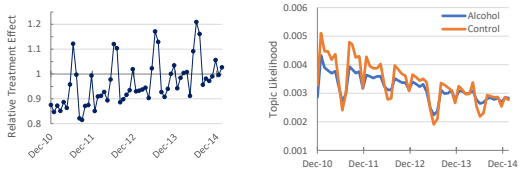
Applying a high-coverage learned mapping from users' profile locations to their geographic regions, we see that 57.2% of our users are placed within the United States, with the largest remainder in the United Kingdom (22.4%) and Canada (2.8%), followed by a broad variety of other countries (14.2%) and unidentifiable or blank profile locations (3.1%) [2]. Note that this skewed distribution to largely English-speaking countries comes about because we restricted our dataset construction to consider only Twitter accounts with an English-language profile. As of Sep. 15, 2010, the users in our dataset had a median of 77 followers (10th percentile=11; 90th percentile=537; max=1.6M). Users followed a median of 104 accounts (10th=23; 90th=485; max=96k). Users had tweeted a median of 1221 tweets prior to Sep. 15, 2010 (10th=60; 90th=10k; max=154k).

Validation of Covariate Balance: To validate that our stratified propensity score analysis is creating statistically comparable alcohol and control groups, we calculate the standardized mean difference (SMD) of each of our covariates in the two groups, as recommended by (Stuart 2010). SMD is defined as the difference in mean covariate value between the two groups divided by the standard deviation of the treatment population. Conventionally, two groups are considered balanced if all covariates have an absolute SMD less than 0.25 (Stuart 2010). Before our analysis, the maximum absolute SMD among our 49k covariates—including tweet frequencies, tweet lengths, and word distributions—was 0.997 standard deviations. After, the maximum absolute SMD is reduced to 0.207, and the median is 0.0124.

Alcohol Mentions

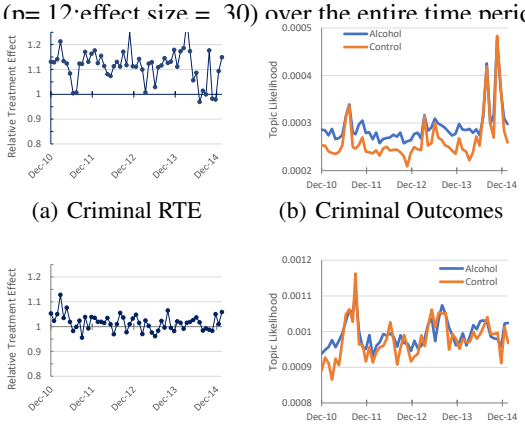
Applying our curated list of alcohol related keywords, we find 4,865,291 tweets from 58,618 unique users over the entire time period and 365,474 tweets from 39,723 unique users

[2] Our geographic inference method achieves high-coverage by mapping profile locations, including colloquial names and non-specific locations ("dmv", "NYC to LHR", "middle of nowhere", "your momma") by empirically learning geographic distributions for each location name from a large, multi-year corpus of geo-tagged tweets (Kıcıman et al. 2014).



(a) Study Habits RTE (b) Study Habits Outcomes

Figure 2: Academic effects: People in the Alcohol group were significantly less likely (p<.05; effect size = .65) to mention studying over the next two years, and somewhat less likely (p= 12;effect size = .30) over the entire time period.



(a) Criminal RTE (b) Criminal Outcomes

(c) Financial RTE (d) Financial Outcomes

Figure 3: Criminal and Financial effects: People in the Alcohol group in the fall of 2010 were more likely to mention legal and criminal challenges through most of our study (p<.001; effect size = .65); and slightly more likely to mention financial pressures over time (p<.1;effect size = .32).

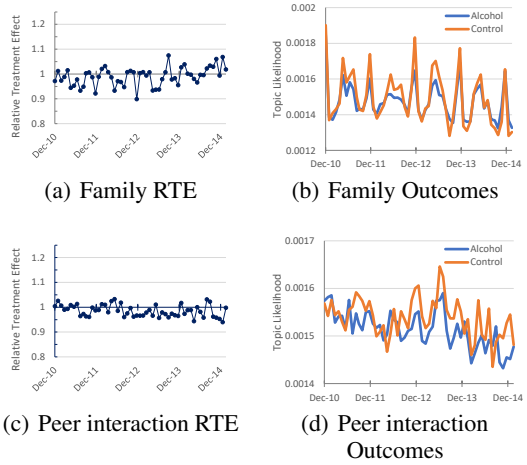
over the Sep. 15 - Dec. 15 2010 treatment window. Table 3 lists the most frequently mentioned alcohol related keywords and examples of tweets matching an alcohol related keyword.

We focus our analysis on the longer-term effects of drinking early in freshmen year. Specifically, we search for alcohol related mentions during Sep. 15-Dec. 15 2010, and measure the effects on word usage from Dec 15 2010 through May 2015. To identify early alcohol drinking, for each user, we calculate the percentage of their active tweeting days during Fall 2010 when they are mentioning alcohol. For example, if a person tweets on 30 days and mentions alcohol on 3 of those days, then we would say that 10% of the person's active days contain alcohol tweets. We define the Alcohol group to be those individuals who are in the top third of the population as ranked by their percent of active days containing alcohol tweets, and the Control group to be those individuals who are in the bottom third. As shown in Figure 1(b), the Alcohol group consists of 21k users who tweeted about alcohol on at least 5% of their active days; and a Control group of 21k users who did not mention alcohol at all in the given window.

Effects of Alcohol Mentions

Effects on College-Success Linked Topics

We quantify the effects of alcohol consumption by first semester college students on a number of topics linked by



(a) Family RTE (b) Family Outcomes

(c) Peer interaction RTE (d) Peer interaction Outcomes

Figure 4: Social effects: People in the Alcohol group were about as likely to mention family (p=.4; effect size = .16), but less likely to mention peer interactions (p<.01;effect size = .61) as people in the Control group.

prior literature to college success and failure. The results of our analysis are shown in Figures 2 - 4, where we see the relative treatment effect (RTE) of alcohol mentions on our selected topics, as well as the weighted (i.e., comparable) measured outcomes for the Alcohol and Control groups. Measured outcomes are topical word likelihoods for the word category (the percent of words in the word category relative to the total number of words). Note that the y-axis values for the topical word likelihoods are irregular because of scale differences. Relative treatment effects are the ratio of the topical word likelihood for the Control group divided by that for the Alcohol group. Relative treatment effects > 1.0 indicate that people who mentioned alcohol in fall 2010 are more likely to mention the given topic than their counterparts. Effects < 1.0 indicate the opposite.

Starting with academic outcomes (Figure 2), we see a relative treatment effect for the Alcohol Group to mention fewer words related to study habits (e.g., 'study', 'library', and 'homework') throughout nearly the entire college time period, and especially early during college. For almost the first two years of college they are 10% less likely to post about study habits, except during summer months when the two groups are effectively identical.

For non-academic outcomes, perhaps the strongest effect was for the Alcohol group to mention more legal and criminal related words for nearly the duration of college, with a relative treatment effect of about a 10% increase in use of such terms over the Control group (Figure 3). Discussion of financial difficulty was approximately the same across the two groups other than a slight increase by the Alcohol group over the second half of the first year of college.

Finally, social measures of family related posting remain fairly similar between the two groups (Figure 4), though if anything there is a slight trend for the Alcohol group to post at slightly lower levels, especially earlier in college. The relative treatment line for instance, generally remains at or below 1.0 for the first half of college. The Alcohol group was

Table 3: Phrases indicating drinking alcohol; and matches across the entire dataset

Phrase	Matched Tweets	Distinct Users	Example tweet
drunk	1,264,158	16,8742	@username hey! getting drunk is not the answer! but I'll happily drink with you
drinking	707,264	15,5541	shouldn't have started drinking this wine :-)
beer	429,853	10,6178	beer in the fridge n I'm ready to go
wine	411,574	10,5103	having a glass of wine this weekend
drinks	389,973	11,3702	Out, had a few drinks, watched a movie and now to bed
alcohol	299,320	9,9295	I mix alcohol all the time
Total	4,865,291	58,618	

generally less likely to engage in friends related posting.

Taken together, these results suggest a picture of those students more likely to mention alcohol early in college as less focused initially on study habits, with increased legal and criminal concerns.Early financial difficulties, a known stressor on college success, and mildly depressed social interactions relative to controls may also contribute.

Effects on Risky Behaviors

One risky behavior is alcohol consumption itself. Previous research supports the conceptualization of alcohol consumption as varied trajectories throughout college and beyond, where some individuals who engage in risky alcohol consumption will continue to do so or increase their use after early adulthood, while others' alcohol consumption will significantly decline as they go through a "maturing out" process as they take on new adult roles (Schulenberg and Maggs 2002; Dawson et al. 2006). To shed light on this, we study early mentions of alcohol by college students in our dataset and the measured effect on their future alcohol mentions. We find that people who mentioned alcohol during the fall of 2010 do mention alcohol at higher rates than their matched counterparts, for nearly the duration of the college time period, as shown in Table 4.

The relative treatment effect drops over the first two years and then stabilizes at about $1.15x$. The weighted (i.e., comparable) proportion of alcohol mentions in the Alcohol group rises slightly over time, though the Control group rises more and meets the Alcohol group toward the end of college.

Alcohol use is not the only risky behavior college students engage in. Thus we also examined relative treatment effects for empath topics of other known risky behaviors. Recall that empath topics are human-validated lexical word categories (Fast, Chen, and Bernstein 2016). Table 4 also includes the effects of early alcohol mentions on people's mentions of sex and party topics. Like the alcohol topic, the sex topic mentions are elevated compared to the Control group, though it decreases over the time period. Notably, the party topic skews toward the Control group effectively throughout the entire time period, lightly suggesting a tendency toward non-social drinking for the Alcohol group.

Effects on Jobs Outcomes

While we do not have data that bears directly on college outcomes, Table 4(e) shows that those in the Alcohol group are less likely to mention work and jobs throughout college, indicating overall less focus on employment, presumably a common transition for graduating students.

Effects on Social-Emotional

Examining social-emotional topics, Table 4(i) and 4(l) show that social topics like friends, family, and home start with an RTE slightly lower than 1 but increase over time. In fact, for some of those topics the RTE flips by the end of college to a positive RTE for those in the Alcohol group. Expression of negative emotion shows an RTE consistently above 1.0, while positive emotion is consistently below 1.0, indicating the Alcohol group is more likely to use negative emotion words and less likely to use positive emotion words than the Control.

Other Strong Effects

We identified empath topics that yielded the strongest effects for either those in the Alochol or Control groups. To do this, we compared the the values of the two time series for each empath topic to identify topics for which the time series for the Alcohol group was notably above that for the Control group or vice versa. Table 5 shows the 10 topics with the strongest effects for the Alcohol group are very physical in nature and include the alcohol topic. For the Control group we see topics that are emotionally positive and church related. Note that after statistical correction for multiple tests, all effects in Table 5 remain highly significant.

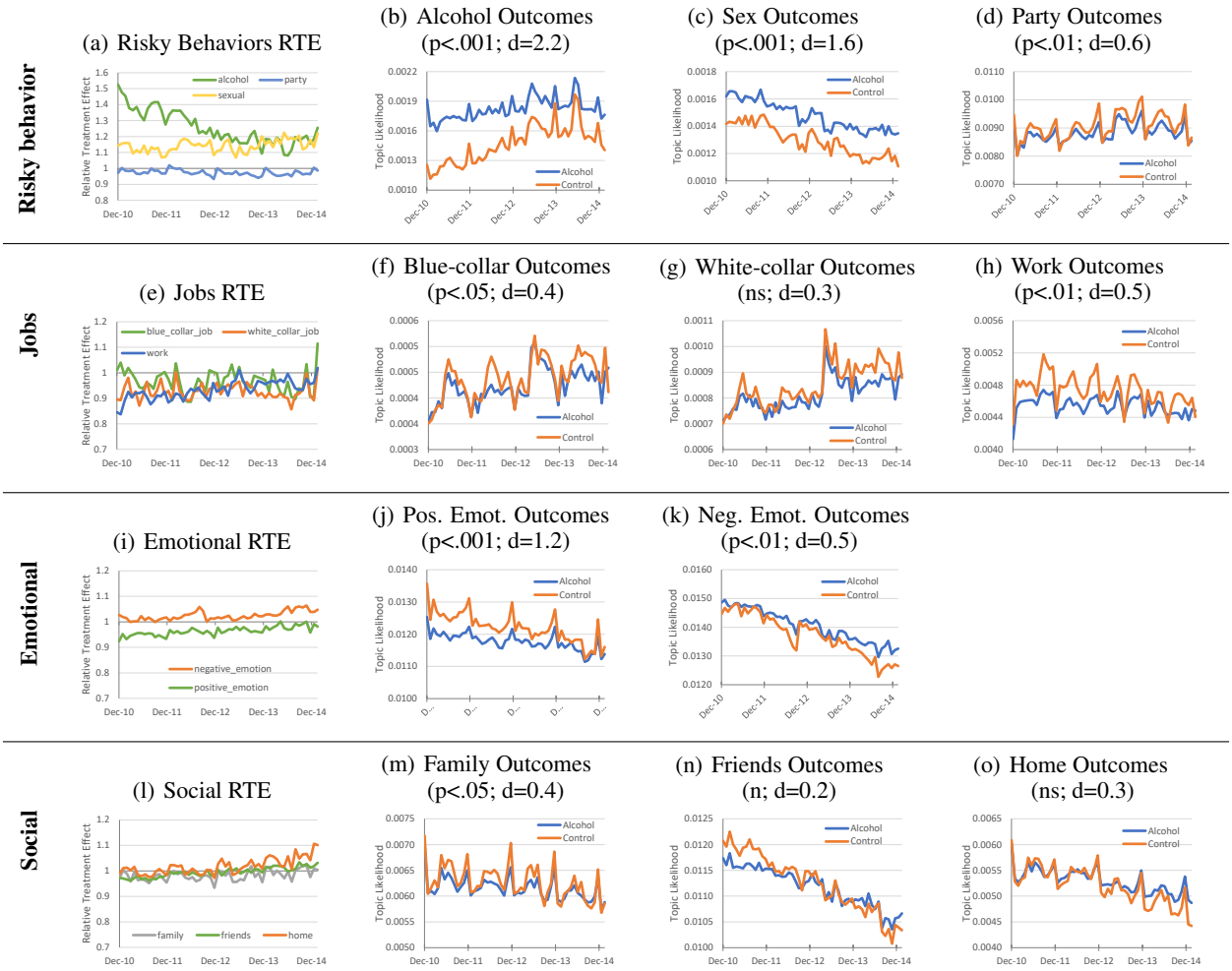
Discussion and Future Work

The propensity score analyses indicate that consumption of alcohol early in college leads to fewer mentions of study habits along with heightened mentions of legal and criminal activity as compared to a stratified sample control group. Further exploration of effects suggests that those who drink early in college are also more likely to mention sexual words, and appear to be less focused on school and work. Social-emotional topics such as positive emotion words and relationship topics like friends and family were lower in the Alcohol group.

This suggests that early drinking, as reflected through social media data, is a correlate of future academic and social red flag indicators of increased student drop out risk. These kinds of analyses, in conjunction with the risk factors identified in prior research, may be able to improve identification of individuals at high risk not only for the current time period, but also help clarify which students will continue engaging in risky alcohol consumption.

One way to leverage these findings is through earlier interventions than those that would be initiated upon poor academic performance, school conduct issues, or legal problems (e.g., an arrest for underage drinking). At the indi-

Table 4: Effects on risky behavior, social-emotional and broader context (p-value, Cohen's d effect size)



Alcohol Topic	Effect	Control Topic	Effect
eating	2.28	worship	-2.80
alcohol	2.19	divine	-2.60
smell	1.70	exasperation	-1.69
liquid	1.69	beauty	-1.66
sexual	1.57	sports	-1.54
cooking	1.51	religion	-1.51
restaurant	1.39	philosophy	-1.47
swearing	1.38	cheerfulness	-1.44
ugliness	1.36	pride	-1.42
hygiene	1.20	optimism	-1.41

Table 5: Empath topics with strongest effects for Alcohol and Control groups. Effect refers to Cohen's d effect size.

vidual level, interventions could include opt-in programs where students receive private feedback via apps or websites on their alcohol exposure. These could be particularly useful during the first semester and year at college when many students are experiencing novel levels of freedom and alcohol exposure. During this transition period, individuals are encountering new norms and establishing new peer groups, and as such, intervention at this stage may be particularly efficacious (Schulenberg and Maggs 2002; Borsari, Murphy, and Barnett 2007). Further, previous research supports a progression model of binge drinkers' behavior of engaging in increasingly more careless and risky behaviors (Vik et al. 2000). Thus early interventions through timely awareness could make critical differences. At the student population level, universities and public health officials can leverage these data to better design programs targeting student awareness of potential consequences of early college drinking, alcohol consumption norms, and corresponding resources to help mitigate these effects.

Methodologically, social media as a data source provides a picture of college student behavior captured in situ, complementing existing methods such as self-report recall studies. Although subject to self-presentation biases, our data show that many social media users in college are sufficiently free in their posts as to make public mentions of alcohol consumption. Previous research similarly suggests that posting alcohol-related content, particularly content which suggest personal drinking, predicts recent drink-

ing, alcohol consumption levels, alcohol use disorder risk, and problems related to drinking (Westgate et al. 2014; Moreno et al. 2015).

One advantage of these data is that consequences of behaviors can be empirically determined from the data themselves rather than predetermined by researchers creating questionnaires and interview protocols. In our example analyses, we focused only on the first semester of college, but similar efforts such as propensity score analyses could be applied to any time frame to examine behavioral outcomes ranging from immediate (e.g., next day) to long term (e.g., college dropout likelihood). For many individuals, binge drinking and alcohol consumption in general decreases after college and their early twenties, but research such as this could help identify which of these risky drinkers are more likely to reduce their drinking after graduating and which are at higher risk of continuing or even increasing their drinking rates (Johnston, O'Malley, and Bachman 1999).

As future work, we plan to extend our study through to 2017 to further examine post-college experiences for those who started in 2010. This will enable analyses such as emotional, relationship, and general life satisfaction coming out of college, as well as any employment and financial outcomes. In addition, repeating our analyses for students beginning college in 2011, 2012 and 2013 will provide a test of the generalizability of our findings. We also plan to more deeply explore the possible heterogeneity of treatment effects, to better understand the specific contexts under which alcohol may have particularly detrimental effects on students.

Conclusion

We examined a dataset of social media timelines of young people entering the transformative life stage of college. For many, college marks a leap forward in independence, responsibility, and behavioral exploration. Prior research highlights the critical nature of college as a step toward future success in life, and yet 40% of students do not finish. Our hope is that detailed longitudinal data such as these can help us empirically explore factors that are predictive of success or failure in key outcomes, particularly so that undesirable outcomes like alcohol abuse and college dropout can be discovered and mitigated in a timely fashion.

Ethical Considerations: Our analyses include potentially sensitive topics, though our use of historical, publicly posted data simplifies some ethical considerations. At no point in our analysis did we attempt to ascertain the real identities of individuals in our dataset. We paraphrased for anonymity the tweets we give as examples. Thus no individual person was identified or could be identified from this work.

Errata

In its original publication, the RTE charts in Table 4 (a,c,i,l) were incorrectly drawn based on older data and did not match the Outcomes charts (b-d, f-h, j-k, m-o). The changes were inconsequential for (a,c,i). Changes to (l) required corrections to the description of *Effects on Social-Emotional* outcomes.

References

Abbar, S.; Mejova, Y.; and Weber, I. 2015. You tweet what you eat: Studying food consumption through twitter. In *Proc. of ACM CHI*, 3197–3206.

Althoff, T.; Jindal, P.; and Leskovec, J. 2017. Online actions with offline impact: How online social networks influence online and offline user behavior. In *Proc. of ACM WSDM*, 537–546. ACM.

Arnett, J. J. 2000. Emerging adulthood: A theory of development from the late teens through the twenties. *American psychologist* 55(5):469.

Borsari, B.; Murphy, J. G.; and Barnett, N. P. 2007. Predictors of alcohol use during the first year of college: Implications for prevention. *Addictive behaviors* 32(10):2062–2086.

Boyle, S. C.; LaBrie, J. W.; Froidevaux, N. M.; and Witkovic, Y. D. 2016. Different digital paths to the keg? how exposure to peers' alcohol-related social media content influences drinking among male and female first-year college students. *Addictive behaviors* 57:21–29.

Clark, J. M., and Halpern, D. F. 1993. The million dollar question: Can an intensive learning experience help lowest-quartile students succeed in college? *Journal of Instructional Psychology*.

Dawson, D. A.; Grant, B. F.; Stinson, F. S.; and Chou, P. S. 2006. Maturing out of alcohol dependence: the impact of transitional life events. *Journal of studies on alcohol* 67(2):195–203.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *Proc. of AAAI ICWSM*.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proc. of ACM CHI*, 2098–2110.

DeBerard, M. S.; Spielmans, G. I.; and Julka, D. L. 2004. Predictors of academic achievement and retention among college freshmen: A longitudinal study. *College student journal* 38(1):66.

Diaz, F.; Gamon, M.; Hofman, J. M.; Kıcıman, E.; and Rothschild, D. 2016. Online and social media data as an imperfect continuous panel survey. *PloS one* 11(1):e0145406.

Duggan, M., and Brenner, J. 2013. *The demographics of social media users, 2012*, volume 14. Pew Research Center's Internet & American Life Project Washington, DC.

Eckles, D., and Bakshy, E. 2017. Bias and high-dimensional adjustment in observational studies of peer effects. *ArXiv e-prints*.

Ellison, N. B.; Lampe, C.; Steinfield, C.; and Vitak, J. 2011. With a little help from my friends: How social network sites affect social capital processes. *The networked self: Identity, community and culture on social network sites* 124–146.

Ellison, N. B.; Steinfield, C.; and Lampe, C. 2007. The benefits of facebook âĂIJfriends:âĂİ social capital and college studentsâĂŹ use of online social network sites. *Journal of Computer-Mediated Communication* 12(4):1143–1168.

Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath: Understanding topic signals in large-scale text. In *Proc. of ACM CHI*, 4647–4657. New York, NY, USA: ACM.

Fourney, A.; White, R. W.; and Horvitz, E. 2015. Exploring time-dependent concerns about pregnancy and childbirth from search logs. In *Proc. of the ACM CHI*, 737–746.

Gerdes, H., and Mallinckrodt, B. 1994. Emotional, social, and academic adjustment of college students: A longitudinal study of retention. *Journal of Counseling and Development: JCD* 72(3):281.

Gong, W.; Lim, E.-P.; Zhu, F.; and Cher, P. H. 2016. On unravelling opinions of issue specific-silent users in social media. In *AAAI ICWSM*.

Gray, R.; Vitak, J.; Easton, E. W.; and Ellison, N. B. 2013. Examining social adjustment to college in the age of social media: Factors

influencing successful transitions and persistence. *Computers & Education* 67:193–207.

Griliches, Z., and Mason, W. M. 1972. Education, income, and ability. *Journal of political Economy* 80(3):S74–S103.

Hargittai, E.; Fullerton, L.; Menchen-Trevino, E.; and Thomas, K. Y. 2010. Trust online: Young adults' evaluation of web content. *International journal of communication* 4:27.

Hingson, R. W.; Zha, W.; and Weitzman, E. R. 2009. Magnitude of and trends in alcohol-related mortality and morbidity among us college students ages 18-24, 1998-2005. *Journal of Studies on Alcohol and Drugs, Supplement* (16):12–20.

Hingson, R. W. 2010. Magnitude and prevention of college drinking and related problems. *Alcohol Research & Health* 33(1-2):45–55.

Imbens, G. W., and Rubin, D. B. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Johnston, L. D.; O'malley, P. M.; Bachman, J. G.; and Schulenberg, J. E. 2011. Monitoring the future national survey results on drug use, 1975-2010. volume II, college students & adults ages 19-50. *Institute for Social Research*.

Johnston, L. D.; O'Malley, P. M.; and Bachman, J. G. 1999. *National Survey Results on Drug Use from the Monitoring the Future Study, 1975-1998. Volume I: Secondary School Students.* ERIC.

Kıcıman, E., and Richardson, M. 2015. Towards decision support and goal achievement: Identifying action-outcome relationships from social media. In *Proc. ACM KDD*, 547–556.

Kıcıman, E.; Counts, S.; Gamon, M.; De Choudhury, M.; and Thiesson, B. 2014. Discussion graphs: Putting social media analysis in context. In *AAAI ICWSM*.

Kıcıman, E. 2012. Omg, i have to tweet that! a study of factors that influence tweet rates. In *AAAI ICWSM*.

Krueger, A. B., and Lindahl, M. 2000. Education for growth: Why and for whom? Technical report, National Bureau of Economic Research.

Lecompte, D.; Kaufman, L.; and Rousseeuw, P. 1983. Search for the relationship between interrupted university attendance. *Acta psychiatt. belg* 83:609–617.

Lu, L. 1994. University transition: major and minor life stressors, personality characteristics and mental health. *Psychological medicine* 24(01):81–87.

Malik, M. M., and Pfeffer, J. 2016. Identifying platform effects in social media data. In *AAAI ICWSM*.

Marczinski, C. A.; Hertzenberg, H.; Goddard, P.; Maloney, S. F.; Stamates, A. L.; and O'Connor, K. 2016. Alcohol-related facebook activity predicts alcohol use patterns in college students. *Addiction Research & Theory* 1–8.

McKenzie, K., and Schweitzer, R. 2001. Who succeeds at university? factors predicting academic performance in first year australian university students. *Higher education research & development* 20(1):21–33.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. N. 2011. Understanding the demographics of twitter users. *ICWSM* 11:5th.

Moreno, M. A.; Cox, E. D.; Young, H. N.; and Haaland, W. 2015. Underage college students' alcohol displays on facebook and real-time alcohol behaviors. *Journal of Adolescent Health* 56(6):646–651.

Musgrave-Marquart, D.; Bromley, S. P.; and Dalley, M. B. 1997. Personality, academic attribution, and substance use as predictors of academic achievement in college students. *Journal of Social Behavior and Personality* 12(2):501.

Myslín, M.; Zhu, S.-H.; Chapman, W.; and Conway, M. 2013. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research* 15(8).

Nelson, T. F.; Xuan, Z.; Lee, H.; Weitzman, E. R.; and Wechsler, H. 2009. Persistence of heavy drinking and ensuing consequences at heavy drinking colleges. *Journal of Studies on Alcohol and Drugs* 70(5):726–734.

Olteanu, A.; Varol, O.; and Kıcıman, E. 2017. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proc. of CSCW 2017*, 370–386. ACM.

Pantages, T. J., and Creedon, C. F. 1978. Studies of college attrition: 1950-1975. *Review of educational research* 48(1):49–101.

Paul, M. J., and Dredze, M. 2011. You are what you tweet: Analyzing twitter for public health. In *Proc of. AAAI ICWSM*, 265–272.

Paul, M. J.; White, R. W.; and Horvitz, E. 2015. Diagnoses, decisions, and outcomes: Web search as decision support for cancer. In *Proc. of WWW*, 831–841. ACM.

Porter, O. F. 1989. Undergraduate completion and persistence at four-year colleges and universities: Completers, persisters, stopouts, and dropouts.

Rosenbaum, P. R., and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Rubin, D. B. 2011. Causal inference using potential outcomes. *Journal of the American Statistical Association*.

Schulenberg, J. E., and Maggs, J. L. 2002. A developmental perspective on alcohol use and heavy drinking during adolescence and the transition to young adulthood. *Journal of Studies on Alcohol, Supplement* (14):54–70.

Schulenberg, J.; Maggs, J. L.; Long, S. W.; Sher, K. J.; Gotham, H. J.; Baer, J. S.; Kivlahan, D. R.; Alan Marlatt, G.; and Zucker, R. A. 2001. The problem of college drinking: Insights from a developmental perspective. *Alcoholism: Clinical and Experimental Research* 25(3):473–477.

Shapiro, D.; Dundar, A.; Wakhungu, P. K.; Yuan, X.; Nathan, A.; and Hwang, Y. 2015. Completing college: A national view of student attainment rates–fall 2009 cohort. *Signature Report* 10.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25(1):1.

Terenzini, P. T., and Pascarella, E. T. 1978. The relation of students' precollege characteristics and freshman year experience to voluntary attrition. *Research in Higher Education* 9(4):347–366.

Tinto, V. 1987. *Leaving college: Rethinking the causes and cures of student attrition.* ERIC.

Valentine, J. C.; Hirschy, A. S.; Bremer, C. D.; Novillo, W.; Castellano, M.; and Banister, A. 2009. Systematic reviews of research: Postsecondary transitions. identifying effective models and practices. *National Research Center for Career and Technical Education*.

Vik, P. W.; Carrello, P.; Tate, S. R.; and Field, C. 2000. Progression of consequences among heavy-drinking college students. *Psychology of Addictive Behaviors* 14(2):91.

Wechsler, H.; Moeykens, B.; Davenport, A.; Castillo, S.; and Hansen, J. 1995. The adverse impact of heavy episodic drinkers on other college students. *Journal of studies on alcohol* 56(6):628–634.

Westgate, E. C.; Neighbors, C.; Heppner, H.; Jahn, S.; and Lindgren, K. P. 2014. I will take a shot for every 'like' i get on this status: Posting alcohol-related facebook content is linked to drinking outcomes. *Journal of Studies on Alcohol and Drugs* 75(3):390–398.

Wolfe, R. N., and Johnson, S. D. 1995. Personality as a predictor of college performance. *Educational and psychological measurement* 55(2):177–185.