

Show me your friends, and I will tell you whom you vote for: Predicting voting behavior in social networks

Lihl Idan

Department of Computer Science
Yale University
lihi.idan@yale.edu

Joan Feigenbaum

Department of Computer Science
Yale University
joan.feigenbaum@yale.edu

Abstract—Increasing use of social media in campaigns raises the question of whether one can predict the voting behavior of social-network users who do not disclose their political preferences in their online profiles. Prior work on this task only considered users who generate politically oriented content or voluntarily disclose their political preferences online. We avoid this bias by using a novel Bayesian-network model that combines demographic, behavioral, and social features; we apply this novel approach to the 2016 U.S. Presidential election. Our model is highly extensible and facilitates the use of incomplete datasets. Furthermore, our work is the first to apply a semi-supervised approach for this task: Using the EM algorithm, we combine labeled survey data with unlabeled Facebook data, thus obtaining larger datasets as well as addressing self-selection bias.

Index Terms—Bayesian Networks, Social Media, U.S. Elections

I. INTRODUCTION

Political choices and voting decisions are considered by many people to be highly sensitive and private information that they are reluctant to reveal. Political campaigns, on the other hand, are investing heavily in voter targeting, focusing intently on social network platforms because of the micro-targeted advertising capabilities they provide. In this paper, we explore the following question: Can we predict the voting behavior of Facebook users from their public Facebook profiles?

We present a novel approach for predicting the voting behavior of Facebook users using a Bayesian-network model that combines demographic, behavioral and social features. Although we focus on the 2016 U.S. Presidential election, our approach can be extended to any two-part system.

This paper is the first to use a Bayesian network model for political attribute inference. Furthermore, it is the first to not only address, but also offer concrete solutions to the selection bias problem, combining a representative and heterogeneous datasets of both politically active and passive users, a semi-supervised training methodology and a diverse

set of demographic, behavioral and social features. Finally, our model is trained to predict not only to *whom* the user will vote, but also *whether* she will vote at all, a task that has not been performed by any existing work on vote prediction.

A. Related work

There are several streams of research that investigate the predictive power of social media for political purposes.

One stream of research concentrates on predicting the political orientation of an individual from various components of her social network profile: Tweets' content [21]; retweet graph [8]; following behavior [3]; degree of tweets and retweets [5], [25]; and "liking" politically oriented pages [4].

A more general approach, taken by [19], [26]–[28] aims at building a generic framework for latent attribute inference of social media users. Those works do not focus on political orientation as a stand-alone trait, but rather use it to demonstrate the functionality of their generic classification system.

Another stream of research focuses on election prediction from social network data. Such works usually take either a volume based approach or a sentiment analysis based approach and use it to predict the outcomes of various election systems in both two-party and multiparty settings [6], [15], [22], [23].

Although closely related to the lines of research discussed above, individual voting behavior differs in several respects. Individual voting behavior prediction aims at predicting an individual choice (unlike election prediction, that aims at capturing an aggregated measure) for a specific event. This event has a well-defined end date and a well-defined set of choices (labels) that represent the possible voting choices the individual has in a given election. In contrast, political orientation reflects a generic and subjective measure that does not have well defined time boundaries; furthermore, the multitude of scales used for measuring political orientation, combined with the fact that any point on such scale may have different meanings to different individuals results in the lack of a well-defined set of labels.

In contrast to the work on political orientation and election prediction, relatively little work has focused on individual voting behavior prediction: Gayo-Avello [9] tried to infer the votes of Twitter users in the 2008 U.S. elections by applying multiple sentiment analysis methods. Bachhuber *et al.* [2] examined several approaches aimed at finding distinct clusters

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ASONAM '19, August 27-30, 2019, Vancouver, Canada

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6868-1/19/08/\$15.00

<https://doi.org/10.1145/3341161.3343676>

of Twitter users based on linguistic properties of their tweets, creating language profiles for supporter groups in the 2016 US elections. Kristensen *et al.* [14] used Facebook to predict voter intentions in the 2015 Danish election, and examined users' political like history to predict which party will they vote for.

B. Shortcomings of prior research

A significant shortcoming of prior research is the use of biased datasets. Those datasets are composed of politically active social-network users, a minority which does not represent the ordinary user population. Classifiers trained on such datasets will thus experience limited predictive accuracy when applied to ordinary users [7], who are less politically engaged yet constitute the majority of social media users [18].

We recognize three potential sources of bias:

Platform: Twitter is the social network most commonly used for politically oriented data-mining research. However, Twitter is one of the least representative social networks [1], [17]: First, the usage statistics among American adults are quite low. Second, Twitter users are not demographically representative of the population; the resulting demographic bias is often ignored in research concerning political orientation inference and voting behavior prediction. Third, Twitter is considered the most "political" social network, attracting a user population with unusually high political awareness. Twitter users' datasets are therefore highly likely to be politically and demographically unrepresentative of the general population.

Features: The vast majority of prior work relies solely on the analysis of user-generated content or politically oriented activities in social networks. This approach introduces substantial selection bias because the only social network users who appear in such datasets are those who engage in politically overt activities, and only a minority of all users do so [7], [14]. In other words, those approaches yield high-accuracy results, but those results are limited to a small fraction of social media users. Indeed, Table II shows that less than 27% of the users in our dataset performed a public activity related to Trump or Clinton before the 2016 election date; that is, more than 73% of the users for whom solely relying on user-generated content would yield results which are essentially no better than those we would have received using a random classifier.

Labels: Previous works have used several methodologies for extracting labels from social-network accounts, relying on online behaviors such as explicitly stating political orientation online [9], [19], [27]; including politically related content in tweets [2], [8], [25]; supporting partisan causes [21] or following candidates [24]. Subsequently, only users for whom such label exists are included in the datasets, leading to the creation of biased datasets, composed entirely of individuals who voluntarily disclose their political preference online. This methodology introduces self-selection bias into the final results, as users who choose to disclose their political affiliations constitute a minority of social media users [20].

C. Our contribution

The goal of the models presented in this paper is the following: given a Facebook user, predict the individual voting

behavior of that user based on the public portions of the user's Facebook profile. Our main contributions are the following:

Addressing sources of potential bias: we address platform-based bias by using Facebook as our social network platform. Facebook is known to be much more representative of the general population than Twitter and supports a richer profile representation; we address features-based bias by going beyond active content analysis and combining multiple types of information about the user: static (demographic attributes of the user), dynamic (activities performed by the user and their frequency patterns) and social (information we can learn about the user from her social network links). Finally, we address label-based bias by applying a semi-supervised approach that uses a combination of labeled data, where labels are obtained from surveys, and unlabeled data, composed of users who didn't participate in the survey though were offered to.

Incorporating non-voters: Existing works implicitly assume that all users indeed vote. That is, the only users that are included in the datasets are users who voted for one of the candidates in a given election. This assumption is not only an unrealistic oversimplification of election systems but also prevents us from identifying important subpopulations, whose specific vote is inconsistent with their general partisanship such as strong partisans who decide not to vote in a given election. Our key assumption is that a voting decision is influenced by two types of components: a "static" component that is determined by the general party identification of the individual, and "dynamic" components that are determined by specific characteristics of the candidates and include both factual policy-related material and subjective perception of the candidate. Each component can be learnt from different elements in a social-network profile, and its influence on the individual's voting behavior is combined with the rest of the components using a Bayesian network model.

Using a novel Bayesian network model: Bayesian-network (BN) classifiers offer important advantages that specifically fit the nature of our problem as well as the nature of our data.

One advantage of BNs is their ability to support the combination of data and prior knowledge about the problem's domain. This allows us to use existing research and statistics for encoding some of the model's parameters; these include interactions between demographic attributes, and interactions between demographic attributes and party identification. Another advantage of BNs is that they handle missing data very well within both training and evidence data. This is particularly important when dealing with social network datasets, which are often incomplete. Indeed, datasets used in this work contain a large number of missing values, which correspond to attributes that users have chosen not to include in their public Facebook profile. Due to our use of BN, missing values in the evidence data need not be imputed but rather can be fed directly to the model. Furthermore, the probabilistic representation combines naturally with the Expectation-Maximization (EM) algorithm, enabling our training data to include both missing values and latent variables. It is this trait of BNs that facilitates the use of both labeled and unlabeled training data.

II. METHODOLOGY AND DATASETS

We designed and distributed a comprehensive survey which adopted purposeful sampling of eligible voters in the 2016 U.S. elections that are also Facebook users. We used the Qualtrics survey platform to create and host the survey.

Survey construction: The survey included questions about the user’s demographics, her Facebook activity, her Facebook friends’ demographics and activity, and her political opinions including her party identification and her vote in the 2016 presidential election. All survey data was anonymized after collection. We informed participants that their responses would be used for academic research.

We implemented several methods for identifying and excluding data from participants who answered unreliably. First, we eliminated responses from participants who took the survey more than once. We took a conservative approach and discarded responses that came from the same IP address. Second, to ensure the eligibility of participants, they had to complete a screening questionnaire before taking the survey. The questionnaire was carefully designed, in order to prevent respondents from inferring the qualifications we were looking for and taking the survey without being an eligible participant. We avoided yes/no questions regarding the qualification needed and used multiway questions instead, as yes/no questions tend to insinuate the “correct” answer in order to pass the screening questionnaire. In addition, we disguised the real screening questions among other dummy questions. For example: instead of directly asking: “do you have a Facebook account”? we asked a series of identical multiway questions about news consumption habits from each of the media platforms. For example, the question “how often do you consume news via TV?” had the following five answers: More than 5 times a day; 3 times a day; once daily; Never; I do not have a TV. In the same question that dealt with Facebook, the last answer was replaced by “I do not have a Facebook account”. Third, we included control questions to ensure that the respondents were providing reliable data; those were fairly straightforward questions which asked the same question multiple times, in different parts of the survey, and using a slightly different terminology. We excluded participants who failed in one or more of the control questions. Lastly, the survey’s name did not include words such as ‘politics’ or ‘social media’ so as to not expose the qualifications needed, as well as not to oversample from the more politically engaged population.

Datasets: In this paper, we make use of three datasets:

\mathcal{D}_1 consists of 1638 survey responses collected via Amazon’s Mechanical Turk (MTurk). Labels for this dataset are obtained through the survey.

\mathcal{D}_2 consists of 841 Facebook profiles and corresponding survey responses collected via both Facebook and Qualtrics. Labels for this dataset are obtained through the survey.

\mathcal{D}_3 consists of 500 Facebook profiles collected via Facebook. Those individuals did not participate in our survey, though were offered to. The only information that was collected on this dataset is *public* — details that the users have

TABLE I: Descriptive statistics for \mathcal{D}_1 and \mathcal{D}_2 . \mathcal{D}_1 was stratified by Party Identification, resulting in a balanced representation of Republicans, Democrats, and Independents within the training set.

Attribute	Metric	\mathcal{D}_2	\mathcal{D}_1	Census
Gender	Female	48.6%	50.6%	50.8%
	Male	51.4%	49.4%	49.2%
Education Level	Post graduate degree	52%	54.5%	62.6%
	College degree	35%	31.2%	26.2%
	No college degree	13%	14.3%	11%
Marital Status	Married	48.9%	49.5%	48.3%
	Unmarried	51.1%	50.5%	51.7%
Race	Caucasian	66.6%	71.4%	62.8%
	African American	10.7%	9.77%	12.2%
	Hispanic	15.6%	12.1%	16.9%
	Other	7.13%	6.72%	8.1%
State Of Residence	Solid republican	28.3%	23.6%	24%
	Lean republican	8.56%	10.5%	11.7%
	Competitive	38%	39.2%	40.5%
	Lean democratic	13.1%	14.6%	14.8%
	Solid democratic	12%	11.9%	9%
Age	20-33 years	38.9%	53.7%	19.7%
	34-49	30.2%	26%	20.5%
	50+ years	30.9%	20.3%	44%
Income	<\$25k	36.9%	38.5%	38.9%
	\$25k-\$60k	47.6%	46.8%	42.2%
	\$60k+	15.5%	14.8%	18.9%
Party Identification	Republican	33.9%	33.3%	
	Democrat	40%	33.3%	
	Independent	26.2%	33.3%	
Vote	Trump	30.8%	32.1%	
	Clinton	36.7%	37.3%	
	Other	32.5%	30.6%	

published in their public Facebook profile. For that reason, \mathcal{D}_3 is unlabeled and contains a large number of missing features.

The use of \mathcal{D}_1 , \mathcal{D}_2 , \mathcal{D}_3 was motivated by two considerations:

Obtaining larger datasets Recruiting a large number of respondents via MTurk (\mathcal{D}_1) was a relatively quick and easy task in comparison to a direct Facebook recruitment process (\mathcal{D}_2). \mathcal{D}_1 , however, was not fully representative of our target population, as it is limited to individuals who are registered to MTurk. Furthermore, tasks published on MTurk cannot ask for personal data, such as name or Facebook profile. In order to account for those limitations, we used \mathcal{D}_1 as our training dataset and \mathcal{D}_2 as our test dataset. A second strategy for obtaining larger datasets is the use of unlabeled data, \mathcal{D}_3 , for augmenting our training data. Such data can be easily collected as the inclusion of a Facebook account in an unlabeled dataset does not require the account holder to participate in the survey.

Addressing selection bias we consider three types of bias:

Visibility bias: Datasets used in prior research were often created by systematically excluding users who did not publish a certain attribute — one that is crucial to the model’s performance, such as politically oriented activity or self-reported label — in their public social network profile. Acknowledging that such approach yield datasets that are not representative of the Facebook user population, our datasets were constructed so that the absence of a feature or a label from a user’s public profile does not affect the user’s chance of being selected; as is evident from Table II, neither feature was publicly disclosed by more than 46% of the users in \mathcal{D}_2 , and only 4.4% of them published their vote in the 2016 presidential election.

Self-selection bias: \mathcal{D}_1 represents a population of individuals

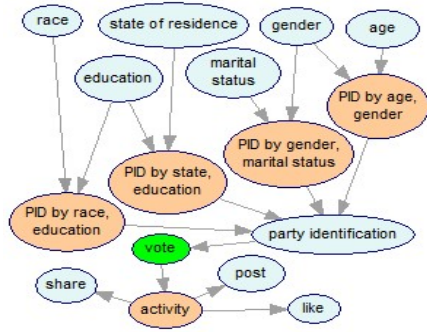


Fig. 1: The BASIC model

who have actively picked our survey from the list of tasks published on MTurk. \mathcal{D}_2 represent a population that was invited by us to participate in the survey and agreed to do so. Those two groups *selected* to participate in our survey. In order to address this selection bias, we created a third dataset, \mathcal{D}_3 , composed of Facebook users who, though were asked to, did not participate in our survey. While we could not use \mathcal{D}_3 for testing, as it is unlabeled, we did combine it in our training set in order to further generalize our model.

Demographic bias: one advantage of the use of MTurk for data collection is the ability to obtain data from demographically diverse groups; although MTurk’s population is not fully representative of the US population, the ease of data collection via MTurk enabled us to reach crowds that are diverse across the primary demographic dimensions used in this work. Table I presents descriptive statistics of \mathcal{D}_1 and \mathcal{D}_2 and compares a subset of their demographics to the 2014 US census. As can be seen, each group is well represented in the datasets.

External software: Throughout this paper we refer to specific subtasks that were performed using existing, readily available software. Examples include Genderize.io¹ and ethnicolr² for gender and race inference based on the user’s name; Python NLTK for sentiment analysis tasks; GeNle and pysmile for our BNs’ creation, parameter learning, and inference.

III. THE BAYESIAN NETWORK

The target node in our BN, “vote”, represents the voting behavior of an individual user. It may take one of three values: vote for candidate 0 (“c0”); vote for candidate 1 (“c1”); not vote at all or vote for a third party candidate (“Other”). In our setting, “candidate 0” represents a vote for Donald Trump; “candidate 1” represents a vote for Hillary Clinton; “Other” represents non voters or a vote for a third party candidate in the 2016 U.S. elections. The basic idea behind our BN models is to treat the vote node as both a cause and an effect. As such, it is influenced by a set of causes and causes a set of effects: Causes include the party identification (PID) of the user and the voting behavior of the user’s network neighborhood (NET). Effects include the user’s politically oriented activities on Facebook (ACT). Each of the causes and effects belongs to a subnetwork that includes

a different type of observable variables: demographic attributes of the user (“static subnetwork”); types and patterns of Facebook activity performed by the user (“dynamic subnetwork”); and activities performed by the user’s network neighborhood (“social subnetwork”). Due to the BN’s structure, we are able to elicit different priors with varying levels of confidence to the different subnetworks.

Formally, given a target Facebook user u_t , Let $D = \{D_t^1, D_t^2, \dots, D_t^n\}$ be n demographic attributes of u_t , $A = \{A_t^1, A_t^2, \dots, A_t^k\}$ be k attributes describing Facebook activities performed by u_t and $N = \{N_t^1, N_t^2, \dots, N_t^j\}$ be j attributes describing u_t ’s network neighborhood. Given our target node, vote, we are interested in the label c that maximizes the following posterior probability:

$$c = \underset{v \in \{c0, c1, Other\}}{\operatorname{argmax}} P(\text{vote} = v \mid PID, ACT, NET) \quad (1)$$

Assuming that $\{D_t^1, D_t^2, \dots, D_t^n\}$, $\{A_t^1, A_t^2, \dots, A_t^k\}$ and $\{N_t^1, N_t^2, \dots, N_t^j\}$ are observable attributes, and using a bayesian approach,

$$P(ACT \mid A_t^1, A_t^2, \dots, A_t^k), P(NET \mid N_t^1, N_t^2, \dots, N_t^j)$$

are obtained from our datasets and a uniform prior, and

$$P(PID \mid D_t^1, D_t^2, \dots, D_t^n)$$

is obtained from our datasets and survey-based, or census-based priors.

Using the EM algorithm, we employed a two-stage learning methodology: first, we trained a BN using the available labeled data (\mathcal{D}_1) and probabilistically labeled the unlabeled data; then, we trained a second BN using both the labeled data and a subset of the unlabeled data (\mathcal{D}_3) about which the first BN was the most confident. This allowed us to both learn the parameters of the model’s latent variables and diversify our training set with a more representative, but unlabeled, data. Inference was done using a Junction tree based algorithm.

Complexity-accuracy trade-off: the models presented in this paper were carefully designed so as to create compact models without damaging performance; if a certain edge did not substantially contribute to the model’s predictive performance, we did not include it in the network. For example, the income node in our SELF INFER model is, theoretically, influenced by many demographic attributes. However, we have found that the only edges that significantly contribute to the overall accuracy were those connecting income with occupation, race and gender, and hence included only those edges in the network.

BASIC: Two nodes, “party identification” and “activity” are directly linked to the “vote” node. “Party identification” covers the influence of an individual’s general party identification on her actual vote. It is considered an unobservable attribute, and thus inferred from the user’s demographic attributes. The attributes which are included in the model met two criteria: attributes which are highly indicative of party identification; observable attributes, in the sense that they are either offered by Facebook as an optional field in a profile or can be inferred from other public information, such as name. The following

¹<https://genderize.io/>

²<https://ethnicolr.readthedocs.io/>

attributes were included in our model: gender, age, race, state of residence, education level, marital status.

We enhance the static subnetwork using prior information about the magnitude of the interactions between each demographic attribute and the PID. Priors are based on multiple surveys conducted by Pew research center³ and on statistics provided by the U.S. Census Bureau; priors refer to data that was collected at least six months before the elections and were incorporated into the model’s parameters using a Bayesian approach and a Dirichlet prior with varying confidence factors based on the available prior information on a given parameter.

As can be seen in Fig. 1, the static subnetwork is built in a conceptually “layered” fashion: The first level represents ascribed attributes; the second level represents acquired attributes; the final level contains the PID node. In order to avoid a large CPT for the PID node, we used a parent-divorcing technique, introducing intermediate nodes (colored orange in Fig. 1,2) that serve as the “accumulators” of the PID given the subset of demographic attributes to which they are linked.

The second node which is directly linked to the target node is the “activity” node, which addresses the more dynamic indicators of the user’s voting behavior and is influenced by the public Facebook activities that the user performs.

In BASIC, we included three types of activities: writing a post; sharing an item; and “liking” a page that is positively or negatively associated with one of the candidates. Each of those activities is represented in the model using a separate observable node. The “activity” node aggregates the different combinations of activities into three states that represent the user’s overall activity: activity associated with supporting candidate 0, activity associated with supporting candidate 1 and activity that is not associated with supporting either candidate (such as writing a positive post about both candidates).

REVISED: Our second model, the REVISED model, refines the dynamic subnetwork (Fig. 2) in order to reflect several insights we obtained from reviewing the explanations respondents gave to different questions in the survey:

Positive activity towards a candidate and negative activity towards her opponent are not equivalent: Although positive activities were almost always associated with a vote for the candidate who was the subject of the activity, negative activities towards a candidate were associated either with a vote for the candidate’s opponent, or with non voters.

Passive users versus politically passive users: The respondents who answered that they did not perform any political activities on Facebook were asked to explain why they didn’t. A common explanation was an unwillingness to reveal political opinions on Facebook. It is important to note that those answers differed from the answers of respondents who reported that they rarely use Facebook, use Facebook without performing activities, or are simply not interested in politics. An interesting observation was that most users who said they did not want to expose their political opinions on Facebook identified themselves as either voted for Trump or did not vote at all.

TABLE II: Fraction of users (D_2) whose public profile contains the following attributes

Attribute	Fraction revealed
Age	18.2%
Educational attainment	34.8%
Marital status	36%
Occupation	32.7%
State of residence	45.5%
Wall content (full)	57.7%
Wall content (partial)	72.5%
Friends list	53.6%
Activity (positive or negative) associated with either Clinton or Trump	26.6%
Vote in the 2016 Presidential election	4.4%

Considering other Facebook activities: Respondents pointed out that, although they have not performed an activity associated with one of the main candidates, they did perform an activity that was associated with a third-party candidate.

Based on our observations, we created the REVISED model: First, the “activity” node was replaced with two nodes: “positive activity” and “negative activity”. In addition, the observed “post”, “share”, and “like” nodes are each replaced by two nodes, one for the positive form of the activity and one for the negative form of the activity. Note, that unlike the BASIC model in which there was no distinction between a positive activity towards candidate x and a negative activity towards her opponent (as both were mapped to the same “pro candidate x ” state), here we distinguish between positive and negative activities, each separately influences the vote node.

Second, we added a new binary node, labeled “Other activity”, that represents various online activities that may indicate a vote for a third-party candidate. In the REVISED model, it is set to one only if there is a positive activity associated with a third party candidate in the user’s profile.

Third, we added a binary node, “activity level”, that represents the general activity level of the user on Facebook. We wanted to distinguish those users who are not active on Facebook from those who are not *politically* active on Facebook; our second observation suggested that while the first group is not particularly important for the inference process, the second group is strongly associated with users who voted for Trump (more generally, with the candidate who is considered less “socially acceptable”). We added four questions to the survey: one asked for the subjective estimation, on a ten-point scale, of the respondent’s activity volume on Facebook. The remaining three asked about the number of posts, shares, and likes that the user has recently performed. We aggregated those four measurements into one node, “activity level”, by discretizing each variable with a median split and setting the “activity-level” value to the four variables’ logical AND.

IV. HANDLING INCOMPLETE DATA

The models discussed thus far has considered a simplified setting, in which no evidence data is missing. While social-networking services can potentially obtain such complete information about their users (by accessing profile items in all visibility levels or collecting demographic data when the user creates an account), third-party services can not, because they are limited to the public portion of social-network profiles that

³<https://www.pewresearch.org/>

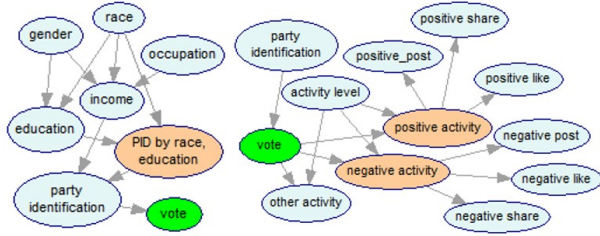


Fig. 2: Additions to the BASIC model of Figure 1 needed to obtain the REVISED (right) and SELF-INFER (left) models

often lacks some of the attributes that comprise our models’ features. Thus, it is essential to understand how well does the model perform in the presence of missing data.

MISSING: In order to understand to what extent our model’s performance is influenced by missing data, we tested the REVISED model on another dataset, $\mathcal{D}'2$, which is identical to $\mathcal{D}2$ except for the following case: if the value of an attribute was not publicly shared in a user’s Facebook profile, we deleted the attribute’s value from the user’s record in $\mathcal{D}'2$ and treated it as “missing evidence”. In other words, $\mathcal{D}'2$ includes only those attributes that the users in $\mathcal{D}2$ have chosen to publish under a “visible to everyone” setting. For attributes that do not have their own field in a Facebook profile, such as race, we used external software and inferred them from other public information such as name; if the software misclassified the attribute (compared to the real label that was obtained from the survey), we deleted it from $\mathcal{D}'2$.

Results for the MISSING model suggested a significant decrease in the overall accuracy, resulting in the need to design models that are specifically suited for missing evidence. The following models demonstrate several strategies that were used to handle the challenge of incomplete datasets.

SELF INFER: This model enriches REVISED by exploring various correlations between the input variables for the purpose of reducing the influence of missing data on the overall results.

Interactions among demographic attributes: Instead of treating demographic attributes uniformly, we partition them into two groups: ascribed attributes and acquired attributes. We can then make use of the possible dependency relations between the two groups. This idea is incorporated in the model by creating new edges between the two “acquired” attributes, education level and marital status, and the “ascribed” attributes which are known to be highly indicative of them: race, gender for the education level, and age, gender for the marital status. An important advantage of this approach is the ability to combine priors in the CPD of the demographic child nodes, which represent the known “influence” of the node’s parents on their descendant; the priors for this purpose were taken from census statistics and were incorporated into the model’s parameters using a Bayesian approach with a Dirichlet prior.

Hidden demographic attributes: Demographic attributes that were included in the former models were both influential on the PID and observable, *i.e.*, can be directly extracted from a Facebook profile. A question that may arise is whether both

of those traits must exist in a single attribute. The key idea is that while some observable attributes are not highly influential on the PID, they are influential on other unobservable attributes which do have a high influence on the PID. Using the chain of observable attribute \rightarrow unobservable attribute \rightarrow PID we can combine such unobservable nodes in our model as well.

We consider one such pair: occupation and personal income. While income is not a “potentially observable” attribute on Facebook, it is highly indicative of the individual’s PID. Occupation, on the other hand, is not considered very indicative of PID but can be considered as an observable attribute. In the model shown in Fig. 2, we incorporate this idea, introducing two new nodes: an observed occupation node and a hidden income node. The latter is fed by three observed nodes: occupation, race and gender, and feeding the PID node. The occupation node’s states represents the twenty main occupation categories according to the standard SOC (Standard Occupational Classification) system, to which we added two additional states: “student” and “retired”.

V. COMBINING LINK INFORMATION: THE FULL MODEL

Homophily is the tendency of individuals to link to others who are similar to them. In a social network setting, this principle implies that the network neighborhood of an individual may reveal a significant amount of information about the individual. In the FULL model, we adopt a fine-grained interpretation of the homophily principle in order to both enrich the static subnetwork and create the social subnetwork.

A common BN representation of social network ties [12], [13] relies on a user-based granularity where each node corresponds to a user within the target user’s (u_t) neighborhood; in such representation, dependency relations (and, consequently, similarity relations) can only be established between *users*. Thus, such representation implicitly assumes that if two users share a tie, all their traits are similar. As demonstrated by [16], this assumption is false: It is known that some attributes experience a more homophilous nature than others. For some, similarity does not induce homophily at all.

In order to facilitate a more realistic representation of social network ties, our FULL model relies on an attribute-based granularity, thus reflecting a more “fine-grained” homophily which may vary across different attributes. To achieve that, we decompose a user into the different attributes that constitute her social network profile by associating nodes with pairs of $\langle \text{user}, \text{attribute} \rangle$, and dependency relations with links between *attributes* of users. Using this fine-grained representation, dependency relations can be established between a single attribute of u_t and the same attribute of each of her neighbors only if the attribute is known to be highly homophilous.

The above idea is implemented in the model using a small, fixed number of “aggregator” nodes, thus avoiding the overhead of dynamically allocating a separate node for each neighbor of u_t . For a given attribute, those nodes aggregate the magnitude of each of the attribute’s states within u_t ’s neighborhood: let a_t be an attribute of u_t that we want to infer from u_t ’s neighbors, and has $q(a_t)$ states, $a_t^1..a_t^{q(a_t)}$. In the BN, we allocate $q(a_t)$

TABLE III: Comparison of the overall classification accuracy using different training configurations.

TEST →	$\mathcal{D}2$			$\mathcal{D}'2$		
TRAIN ↓	BASELINE	BASIC	REVISED	MISSING	SELF INFER	FULL
10-fold CV	.382	.68	.745	.661	.718	.825
$\mathcal{D}1$.382	.726	.76	.67	.712	—
$\mathcal{D}1+\mathcal{D}3$.382	.738	.782	.695	.755	—

aggregator nodes, parents of a_t , where the i th node represent the magnitude of the subset of u_t 's neighbors for whom the value of a_t equals the state a_t^i .

A. Enriching the static subnetwork

We use the homophily principle to infer missing demographic traits of u_t from the Facebook profile of u_t 's neighbors.

FULL: In the FULL model, we focused on a particular subset of ties within u_t 's network neighborhood. The subset is composed of neighbors who are "close friends" of u_t , where a close friend is defined as a friend who has reacted to a recent post on u_t 's wall. Using this approach, we only consider ties that are both intense and recent, as the stronger the tie connecting two individuals, the more similar they are [11], [16]. Furthermore, such activity based metric is more predictive of tie strength than metrics based on network structure or social distance, incorporating both intimacy and intensity factors [10], [11]. The major advantage of this approach is its simplicity and practicality, as it does not require access to the full list of friends, but only to u_t 's wall. Furthermore, such ties can be identified and accessed directly from u_t 's profile.

We examined two attributes that are known to be highly homophilous: state of residence (SOR) and age. We elaborate on the CPD design of the state of residence node (V_{SOR}). The CPD of the age node was built in a similar process.

The influence of u_t 's neighbors' SOR on u_t 's own SOR is modeled using five aggregator nodes that represent the five states that the SOR variable can take in the BN (see Table I), $\{SOR^i\}_{1 \leq i \leq 5}$. The j th aggregator node quantifies the portion of u_t 's close friends for whom $V_{SOR}=SOR^j$. Each of the aggregator nodes was discretized to reflect the rank of the state it represents among V_{SOR} 's states and was linked to V_{SOR} as its parent using a memory-efficient multinomial logistic CPD.

B. Creating the social subnetwork

The influence of the target user's neighborhood on her voting behavior has been considered thus far only indirectly, through the static subnetwork. A natural extension is to treat the vote node as a stand-alone trait; then, the homophily principle can be used directly to infer the voting behavior of u_t from the voting behavior of her neighborhood. However, because the voting behavior cannot be directly extracted from a Facebook profile field we cannot use exact counts as we did for demographic attributes. Instead, we use an approximation of the overall voting behavior of u_t 's neighborhood, obtained via the following procedure: First, we mark the subset of u_t 's neighborhood whose voting behavior can be estimated with high confidence; these are users who have performed a social network activity that is associated with one of the candidates.

Second, we determine the influence of u_t 's neighborhood on her own voting behavior using *only* this marked subset, as well as the magnitude of this subset relative to u_t 's neighborhood.

FULL: A social subnetwork, built under the same principles used to enhance the static subnetwork, is added to the FULL model. The subnetwork contains four aggregator nodes, representing the portion of u_t 's close friends who performed either a positive or negative Facebook activity associated with candidate 0 or 1, and a "noise" node, representing the portion of u_t 's close friends who did not perform any public Facebook activity associated with a candidate. In order to avoid continuous-valued nodes, the portions represented by each node were discretized into five intervals using equal-frequency discretization. All four nodes were linked into one of two intermediate nodes ('Positive neighborhood activity' and 'Negative neighborhood activity'), while the 'noise' node was linked to both. Those intermediate nodes aggregate the influence of the user's neighborhood on her voting behavior and feed directly to the vote node.

VI. EXPERIMENTAL RESULTS

The BASELINE, BASIC and REVISED models were evaluated on $\mathcal{D}2$, which includes attributes in all visibility levels. The REVISED model was also evaluated on $\mathcal{D}'2$, which only includes attributes from $\mathcal{D}2$ published under a "visible to everyone" setting. SELF INFER and FULL, designed to handle missing evidence, were evaluated on $\mathcal{D}'2$.

Table IV provides a detailed summary of results. We report overall accuracy, Precision, Recall, and AUC for each class. For FULL, we report the 10-fold cross-validation results on $\mathcal{D}'2$, because $\mathcal{D}1$ does not include social features. However, in order to take advantage of the information in $\mathcal{D}1$, we use the CPDs obtained for the static and dynamic subnetwork in previous models as high-confidence priors for FULL.

Inspired by [19], we employ a simple baseline system (BASELINE) that classifies all the users explicitly mentioning their vote in the 2016 election within one of their public posts. All other users are considered misses for the given class.

Table III compares the overall accuracy of each model using different training configurations. The cross-validation results allow adequate comparison between the results of FULL in Table IV and the rest of the results. As can be seen, augmenting the labeled dataset, $\mathcal{D}1$, with unlabeled data, $\mathcal{D}3$, indeed improves the classification accuracy. Furthermore, this increase becomes more significant as the number of missing values within the test set increases.

BASELINE's low overall accuracy probably results from the fact that very few users have publicly disclosed their voting intention. Results for the REVISED model demonstrate an improvement over the BASIC model, especially for the "other" class. This improvement suggests that the finer-grained modeling of the dynamic subnetwork allows the classifier to capture additional information which uniquely characterizes the "other" class, thus better separating it from the other two classes. Results also show that the MISSING model underperforms the REVISED model. However, the differences are not as sharp as we expected considering the large number of missing values.

FULL was the best model, as evidenced by its high accuracy score (82.5%). Its use of social features boosts both precision and recall, but the SELF INFER model, which only includes demographic and behavioral features, achieves a decent score (75.5%) as well. This demonstrates the fact that using carefully designed models combined with complementary sources of information about the user yields solid predictions even when both training and testing datasets are incomplete.

Most determinant features: We performed a sensitivity analysis to assess the impact of the various models’ features on our target variable, vote. We elaborate here on the two most determinant features for each state of the vote node. The most influential feature for vote=Trump was race=African American, which decreased the posterior probability of vote=Trump by 19%, followed by positive_neighborhood_activity_Trump=very high, which increased the posterior probability of vote=Trump by 13%. For vote=Clinton, the two most influential features were positive_post=Clinton and education_level=post graduate degree; they increased the posterior probability of vote=Clinton by 22% and 15%, respectively. The two most influential features on vote=Other were negative_activity=both candidates and noise=very high. They increased the posterior probability of vote=Other by 15% and 9%, respectively.

Comparison with Existing Results: Unlike political orientation, prediction of individual voting behavior has not been intensively studied. Exceptions include [2], [9], and [14], achieving an overall accuracy of 63%, 78.8%, 70.8% respectively. However, datasets used in all three works were artificially limited to include only politically active users: users who performed politically oriented tweets [2], hashtags [9] or likes [14]. The last is highlighted because the accuracy of political-orientation classifiers is heavily dependent on users’ political-engagement level. For example, [7] showed that methods for inferring the political orientation of social-network users previously claimed to achieve greater than 90% accuracy on politically active users, actually achieve barely 65% accuracy when applied to “politically modest” users.

The prior work most relevant to ours is that of Kristensen *et al.* [14], which considers five models. We could not use all five as baselines for our work, because some use features that can not be directly extracted from a Facebook profile (such as the user’s opinion towards politically dividing issues, information which [14] obtains from surveys). In contrast, our main principle in this work is to only use features that can be extracted from a Facebook profile directly, without requiring the user to participate actively in the data-collection process (our survey is used for validation purposes only). We were able to apply model 2, which uses the user’s single most recent like that is associated with a party or politician’s page; model 3, which uses the number of such likes over the past two years; and their final model, which combines the features from model 3 with all the features from the survey that served as their baseline model. We selected from their final model only those features that can be extracted from a Facebook profile directly: gender, age, geography and education. All models were implemented

TABLE IV: Detailed results for the Bayesian network models presented in this paper

Model	Class				Acc.
		Prec.	Rec.	AUC	
BASIC	Trump	.77	.76	.91	.738
	Clinton	.72	.89	.91	
	Other	.72	.53	.78	
REVISED	Trump	.73	.85	.94	.782
	Clinton	.79	.83	.92	
	Other	.83	.64	.84	
MISSING	Trump	.63	.84	.88	.695
	Clinton	.75	.7	.85	
	Other	.72	.53	.78	
SELF INFER	Trump	.69	.87	.91	.755
	Clinton	.81	.73	.88	
	Other	.76	.65	.82	
FULL	Trump	.78	.88	.94	.825
	Clinton	.83	.86	.93	
	Other	.85	.72	.91	

using Python’s scikit-learn library.

Applying model 2 to our dataset (Using \mathcal{D}_1 and \mathcal{D}_2 as training and test sets) achieved an overall accuracy of 38.6%, a low score both compared to the result achieved in [14] (43.9%) and to our BN models’ results. Applying model 3 to our dataset achieved an overall accuracy of 40.4%. It is interesting to see that there is not much improvement in the accuracy compared to model 2, while the difference between the two models in [14] is more significant, with model 3 achieving 60.9%. Finally, applying their final model to our dataset achieved an overall accuracy of 60%; we can see a clear increase in the accuracy compared to models 2 and 3; however, it still underperforms [14]’s final results (70.8%). This results from the fact that [14] both artificially limit their datasets to users who performed a “political like” (only 19% from our dataset) and use features that can not be directly extracted from a user’s profile such as opinion about politically dividing issues. As evidenced by Table IV, [14]’s final model significantly underperforms all our BN models when applied to our datasets. This highlights the added value that BNs have for this specific task compared to a regression model, and the importance of the other features that are not included in [14]’s final model but are included in our BN models.

Apart from [14], no other existing work could serve as a proper baseline for our work, primarily because their objective is different from ours; Unlike works such as [15], [22], we are *not* trying to forecast the general outcome of an election. Unlike works such as [7], [8], [19], [27], we are *not* trying to predict a general political orientation of an individual. Therefore, a direct comparison of our results and theirs is meaningless. For election forecasting, such comparison is impossible since the evaluation metrics are different: while we use an overall accuracy score, election prediction papers use MSE and compare their forecasting results to national surveys. Furthermore, the vast majority of political orientation inference papers have dealt with Twitter; thus, their models heavily rely on Twitter-specific features, making it impossible to apply those models

directly to our datasets. However, in order to gain further insights about how our BN compares to other models used in previous work, we chose two other classifiers commonly used in previous work on political orientation inference: Support Vector Machines (SVM) and Boosted Decision Trees (BDT), as well as simplified Multinomial Naive Bayes (MNB) classifier, and tested the performance of each classifier when applied to our datasets and fed with the features of each of our BN models.

Hyperparameters (cost and γ for the RBF SVM; number of trees and learning rate for BDT) were chosen using a randomized grid search with 5-fold cross-validation. All three models were implemented using Python's scikit-learn library. As seen in Fig. 3, both MNB and SVM considerably underperform the BN on all five configurations. On the other hand, the BDT classifier performs on par with the BN; it slightly outperforms the BN on the REVISED model but underperforms the BN on the MISSING, SELF-INFER and FULL models; that is, when the test set contains missing evidence.

VII. CONCLUSION

In this paper, we presented a novel approach to predicting the voting behavior of Facebook users based on a Bayesian-network model that combines diverse yet complementary types of information about the user. In contrast to previous works, we made use of datasets of ordinary Facebook users, thus avoiding the bias entailed in cherry-picked datasets, limited to politically active users. Using a semi-supervised methodology, we applied our model to the case of the 2016 U.S. elections, achieving promising results despite large amounts of missing data.

Interesting avenues for future research include augmenting the model with additional behavioral and interest based traits, combining more complex measures of the network neighborhood, and adding temporal features that capture how various attributes change over time. The latter is particularly useful for politically-oriented applications such as identifying swing voters by examining whether the political opinions of a user are consistent over time, or measuring the extent to which a specific event influenced the user's voting intentions.

ACKNOWLEDGMENT

The first author was supported in part by US National Science Foundation grants CNS-1407454 and CNS-1409599 and William and Flora Hewlett Foundation grant 2016-3834. The second author was supported in part by US National Science Foundation grants CNS-1407454 and CNS-1409599 and US Office of Naval Research grant N00014-18-1-2743.

REFERENCES

- [1] "https://www.pewinternet.org/2018/03/01/social-media-use-in-2018/,"
- [2] J. Bachhuber *et al.*, "A linguistic analysis of US twitter users," in *Designing Networks for Innovation and Improvisation*, 2016.
- [3] P. Barberá, "Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data," *Political Analysis*, vol. 23, 2015.
- [4] R. Bond and S. Messing, "Quantifying social media's political space: Estimating ideology from public revealed preferences on facebook," *American Political Science Review*, vol. 109, pp. 62–78, 02 2015.
- [5] A. Boutet, H. Kim, and E. Yoneki, "What's in your tweets? I know who you supported in the UK 2010 general election," in *ICWSM*, 2012.

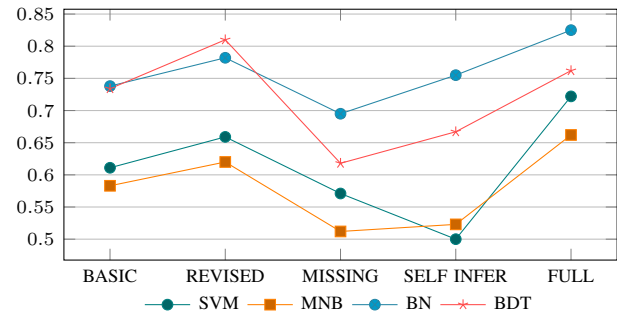


Fig. 3: Overall classification accuracy among different classifiers

- [6] M. C. MacWilliams, "Forecasting congressional elections using facebook data," *PS: Political Science & Politics*, vol. 48, pp. 579–583, 2015.
- [7] R. Cohen and D. Ruths, "Classifying political orientation on twitter: It's not easy!" in *ICWSM*, 2013.
- [8] M. Conover, B. Gonçalves, J. Ratkiewicz, and A. Flammini, "Predicting the political alignment of twitter users," in *IEEE SocialCom*, 2011.
- [9] D. Gayo-Avello, "Don't turn social media into another 'literary digest' poll," *Commun. ACM*, vol. 54, no. 10, pp. 121–128, 2011.
- [10] E. Gilbert and K. Karahalios, "Predicting tie strength with social media," in *CHI*, 2009.
- [11] M. Granovetter, "The strength of weak ties: A network theory revisited," *Sociological Theory*, vol. 1, pp. 201–233, 1983.
- [12] J. He, W. W. Chu, and Z. Liu, "Inferring privacy information from social networks," in *ISI*, 2006.
- [13] J. Jia, B. Wang, and L. Zhang, "Attrinfer: Inferring user attributes in online social networks using markov random fields," in *WWW*, 2017.
- [14] J. Kristensen *et al.*, "Parsimonious data: How a single facebook like predicts voting behavior in multiparty systems," *PLOS ONE*, 2017.
- [15] D. A. Linzer, "Dynamic bayesian forecasting of presidential elections in the states," *Journal of the American Statistical Association*, vol. 108, pp. 124–134, 2013.
- [16] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, 2001.
- [17] A. Mislove, S. Lehmann, Y. Ahn, J. Onnela, and J. N. Rosenquist, "Understanding the demographics of twitter users," in *ICWSM 2011*.
- [18] E. Mustafaraj, S. Finn, C. Whitlock, and P. T. Metaxas, "Vocal minority versus silent majority: Discovering the opinions of the long tail," in *IEEE SocialCom*, 2011.
- [19] M. Pennacchiotti and A. Popescu, "A machine learning approach to twitter user classification," in *ICWSM*, 2011.
- [20] A. Priante, D. Hiemstra, T. van den Broek, and A. Need, "# whoami in 160 characters? classifying social identities based on twitter profile descriptions," in *NLP+CSS*, 2016.
- [21] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *SMUC*, 2010.
- [22] E. A. Stoltenberg, "Bayesian forecasting of election results in multiparty systems," Master's thesis, 2013.
- [23] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welp, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *ICWSM*, 2010.
- [24] S. Volkova, G. Coppersmith, and B. V. Durme, "Inferring user political preferences from streaming communications," in *ACL*, 2014.
- [25] F. M. F. Wong, C. W. Tan, S. Sen, and M. Chiang, "Quantifying political leaning from tweets and retweets," in *ICWSM*, 2013.
- [26] W. Youyou, M. Kosinski, and D. Stillwell, "Computer-based personality judgments are more accurate than those made by humans," *Proceedings of the National Academy of Sciences*, vol. 112, pp. 1036–1040, 2015.
- [27] F. Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors," in *ICWSM*, 2012.
- [28] E. Zheleva and L. Getoor, "To join or not to join: the illusion of privacy in social networks with mixed public and private profiles," *WWW*, 2009.