Expert Systems With Applications xxx (xxxx) xxx

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Learning individual preferences from aggregate data: A genetic algorithm for discovering baskets of television shows with affinities to political and social interests

Balaji Padmanabhan [a,*], Arash Barfar [b]

[a] School of Information Systems and Management, Muma College of Business, University of South Florida, 4202 F. Fowler Avenue, Tampa, FL 33620, USA
[b] Department of Information Systems, College of Business, University of Nevada, Reno, 1664 N Virginia St, Reno, NV 89557, USA

ARTICLE INFO

Keywords:
Genetic algorithms
Multi-objective evolutionary algorithms
Audience analytics
Aggregate data
TV panel data
Micro-targeting

ABSTRACT

This paper presents a flexible general-purpose framework using genetic and multi-objective evolutionary algorithms that can leverage "unlabeled" (and anonymized) panel data on television viewership along with aggregate-level vote or public opinions statistics to (*i*) identify sets of programs that have affinities with politics and social issues, and (*ii*) estimate individual preferences from unlabeled data. The applications of this framework are significant given the wide interest in using big data for political advertising and building election forecasting models with non-polling data. Analyzing viewership spanning over seven billion minutes from Nielsen's TV panel for an entire year (2016), we illustrate how this framework can learn interesting baskets of programs whose viewership can help estimate individual attitudes toward politics, global warming, same-sex marriage, and abortion.

1. Introduction

Digital age notwithstanding, the lion's share of political campaigns expenditure in the United States still goes to television advertising. In the 2018 midterm elections, for example, political advertisements spending hit historic highs, with $4.3 billion on broadcast and cable television and $0.95 billion on digital media (Passwaiter & Meininger, 2018). Political ad spending is now approaching $10 billion in the 2020 presidential election (Bruell, 2019). Thus, a significant part of the campaigns' analytics efforts is devoted to making informed decisions on television advertisements. For example, the proprietary algorithms that helped Obama's reelection in 2012 have been commercialized (e.g., by Civis Analytics) to help "choose the combination of shows that provide the most efficient audience exposure to meet goals."

Broadcast and cable television advertising is addressable at the program level; that is, advertisers can purchase commercial slots across a set of specific programs. Over-the-top media services (which bypass traditional platforms and stream the video content over the Internet), however, allow political campaigns to micro-target at both programs and household levels; an opportunity that comes with a question: what set of programs and/or households should be targeted with an advertisement? This question is not specific to political campaigns; consider a global warming campaign whose goal is to first raise funds for their cause, and then use the funds to raise awareness about climate change. In the former (fundraising) case, the campaign needs to target television programs that have appeal for believers in global warming (i.e., program-level targeting), and/or viewers who are concerned about global warming (i.e., household-level targeting). In the latter (awareness) case, on the other hand, the campaign needs to target programs that have appeal for non-believers in climate change, and/or viewers who disregard global warming.

While individual and household television viewing data might be available through over-the-top media service providers or market data vendors, individual political and social *preferences* are typically unknown. Particularly, it is impractical for media service providers or panel data vendors to ask television viewers potentially sensitive questions about their social or political beliefs. Thus, there exists no "labeled data" (on individual political or social preferences) that can be used to train machine learning models. Motivated by this problem, this paper proposes a framework based on genetic algorithms (GAs) and demonstrates how anonymized panel data on television viewership along with aggregate vote or public opinions statistics can be used to (*i*) identify sets (hereafter "baskets") of episodic programs that have affinities with politics and social issues (e.g., global warming, same-sex marriage, and

* Corresponding author at: 4202 E. Fowler Avenue, CIS 1040, Tampa, FL 33620, USA.
E-mail addresses: bp@usf.edu (B. Padmanabhan), abarfar@unr.edu (A. Barfar).

https://doi.org/10.1016/j.eswa.2020.114184
Received 22 March 2020; Received in revised form 23 September 2020; Accepted 28 October 2020
Available online 3 November 2020

abortion), thereby informing program-level advertising targeting, and (*ii*) estimate individual viewer political and social preferences from unlabeled data, thereby informing household-level advertising targeting. Besides advertising targeting, the proposed framework in part contributes to gauging the political mood, which polls and pundits that rely on them have been clearly struggling with.

We apply the proposed framework to Nielsen's highly granular "television watch data" of 48,816 households throughout the US in 2016, representing over seven billion minutes of television viewership. It is however important to note that our framework is minimally intrusive. In contrast to the Facebook applications of previous campaigns that directly and indirectly solicited participants' political preferences (Tobias, 2018), our framework only uses completely de-identified household TV viewing history. Specifically, the framework requires no demographic data except the household state (e.g., New York). The only political and social data used by the framework are aggregate vote and public opinions data.

The closest empirical work to ours is the Norman Lear Center study by Blakley, Watson-Currie, Shin, Valenti, Saucier, and Boisvert (2019), which examines the relationship between American TV viewing preferences and political/ideological views. There are, however, several distinctions between our work and Blakley et al. (2019). First, the Norman Lear Center study only examines 50 TV shows, whereas our study concerns 600 popular TV programs throughout the US. Second, Blakley et al. (2019) surveyed 3,096 participants about their favorite programs, whereas we use the actual TV viewership data of 48,816 Nielsen panelists throughout a full year. Third, Blakley et al. (2019) conducted their analyses on labeled data (i.e., they asked their participants about their political beliefs), whereas our data is completely unlabeled (we do not know about the individual viewer political/social attitudes). Finally, Blakley et al. (2019) conducted statistical cluster analysis on labeled data, whereas we design and implement a genetic algorithm framework that can learn baskets of shows with affinities to political and social issues, in tandem with individual political/social preferences merely from aggregate vote and opinions statistics.

Methodologically, our work has connections to estimating disaggregate choice models using aggregate data in the marketing literature. Chen and Yang (2007) initiated this research stream through Bayesian analysis of aggregate data where they applied the Metropolis-Hasting algorithm to augmenting individual brand choices from their posterior distribution. Musalem, Bradlow, and Raju (2008, 2009)) then extended the framework by estimating individual brand choices via Gibbs sampling. Our genetic algorithm framework, too, estimates individual political/social preferences from aggregate vote and public opinions statistics. However, in addition to individual political/social attitudes, our framework also learns another variable of interest (i.e., baskets of programs whose viewership can determine the individual's stance on politics or social issues)—thereby extending the disaggregate choice model estimation research. Our work is also related to political recommender systems research. We will discuss these connections further in the paper.

The remainder of this paper is organized as follows. Section 2 presents a brief review of previous research in advertising micro-targeting and political recommender systems. Section 3 explains the television viewership data in detail. Section 4 provides the formalism and problem statement. Section 5 explains our proposed framework based on genetic algorithms. Section 6 provides the results of applying the framework to the US politics and social issues (namely global warming, same-sex marriage, and abortion) and a discussion of the findings. Section 7 presents an extension of the framework using Multi-Objective Evolutionary Algorithms (MOEA) in cases where estimating individual preferences is done with multiple objectives. Section 8 finally presents the conclusions drawn from our work and outlines its limitations.

2. Related work

Our work is related to research in advertising micro-targeting and political recommender systems. This section explores these connections in detail.

2.1. Television advertising targeting

Traditional television viewership is linear; that is, the audience watch the programs at the time they are broadcast by the stations. The linear TV advertising market is typically "upfront," where advertisers commit in advance to purchase commercial slots across a set of shows for a full network season.

Over-the-top media services (which stream the video content over the Internet), set-top boxes, and digital video recorders have made TV advertising addressable. With these platforms, advertisers can micro-target at the household-program level. Household past media consumption data (e.g., TV viewing history), behavioral data (e.g., purchase RFM), and characteristics (e.g., demographics and geography) can potentially facilitate advertising micro-targeting (Malthouse, Maslowska, & Franks, 2018). Previous research demonstrates the benefits of micro-targeting based on predicted households' advertising avoidance (see, for example, Tuchman, Nair, & Gardete, 2018), which in turn could be explained with household-specific characteristics (Deng & Mela, 2018).

The proposed framework in this study contributes to both traditional (linear) and addressable (micro-targeting) TV advertising. To wit, the genetic algorithm discovers baskets of episodic programs with affinities to politics and social issues. Public campaigns can accordingly purchase commercial slots across the discovered programs that match their agenda in the traditional TV advertising market. In addition, the algorithm predicts individual viewer political and social preferences. Political campaigns can subsequently draw on the findings to micro-target households with advertisements that best match the households' political attitudes (for a recent review of empirical research on political advertising, see Lovett, 2019).

It is important to note that different forces are still shaping the future of television advertising. On the one hand, mergers between media and internet companies (e.g., acquisition of NBC Universal by Comcast) facilitate consolidating different data (i.e., past media consumption, purchase behavior, and demographics) that are usually owned by different stakeholders. On the other hand, privacy protection regulations (e.g., GDPR) are addressing consumer privacy concerns raised as micro-targeting practices grow. In this context, the proposed framework in this paper is minimally intrusive; that is, the only household-specific data it uses is the de-identified TV viewing history. The aggregate-level vote and public opinions data used by framework are all publicly available.

2.2. Political recommender systems

In the process of discovering baskets of programs, our framework also estimates viewer political and social preferences—which in turn relates the present study to recommender systems research. As software systems that suggest potentially interesting items to the user based on her predicted preference, recommender systems are applied to political domains as well. As one application in politics, political recommender systems have been used to enhance voter turnout and help citizens choose a candidate that best match their preferences. In Switzerland, for example, Voting Advice Applications (VAAs) provide the voter with a list of recommended parties based on the information provided by both parties and the voter (Meier, 2012).

Another application of recommender systems in politics concerns political marketing (Lovett, 2019). Unlike VAAs in which voters disclose their preferences in the software, here user political and social attitudes need to be estimated by the recommender system for advertising micro-

Table 1
Data.

		600 *episodic programs*					
Household ID	State	*Anderson Cooper*	*Duck Dynasty*	…	*Sean Hannity*	*Basketball Wives*	Political preference(unlabeled)
00001	Florida	0	0	…	1	1	Republican
00002	Texas	0	1	…	0	1	Democrat
…	…	…	…	…	…	…	…
48,816	California	1	0	…	1	0	Republican

targeting. Building models that predict user political attitudes, however, requires labeled data which is not always available.

Researchers have taken two approaches to tackle the challenge of unlabeled data in estimating user political attitudes. The first approach involves surveys. Kosinski, Stillwell, and Graepel (2013), for example, labeled private traits and attributes of nearly 58,000 volunteers through questionnaires in Facebook. While nearly 20,000 volunteers revealed their religions, only 10,000 participants disclosed their political views (i.e., "Liberal" versus "Conservative") in the questionnaires. Youyou, Kosinski, and Stillwell (2015) also used the same data to show that computer models predict people's political attitudes more accurately than their acquaintances.

In addition to being expensive, surveying individuals to label their political attitudes is deemed ineffective given how even exit polls have been failing to predict political outcomes recently. To this end, the second approach to labeling political preferences involves manual coding by annotators. Conover, Goncalves, Ratkiewicz, Flammini, and Menczer (2011), for example, used two annotators to label 1,000 Twitter users for their political preferences.

The goal of this study is to learn the sets of TV programs that have appeal for individuals with Conservative or Liberal attitudes, while panelists are not labeled for their political attitudes—which poses the same challenge of unlabeled data in political recommender systems. In this paper we address this problem in a novel manner by exploiting publicly available aggregate data (e.g., election and public opinion statistics at the state level), which in part relates our work to estimating disaggregate choice models using aggregate data in marketing research. Chen and Yang (2007) first approached this problem through Bayesian analysis of aggregate data where they applied the Metropolis-Hasting algorithm to augmenting individual brand choices from their posterior distribution. Musalem, Bradlow, and Raju (2008, 2009)) then extended Chen and Yang's (2007) framework by estimating individual brand choices via Gibbs sampling. Notwithstanding the similarity, it is important to note the differences between the marketing studies and our work. In the marketing studies, the (aggregate) brand market shares are known while the individual brand choices are not and need to be estimated. In our work, on the other hand, the (disaggregate) individual TV viewership and the aggregate vote shares are known, but the individual political preferences are not known and to be estimated. However, in addition to individual political attitudes, our framework also learns another variable of interest (i.e., baskets of programs whose viewership can determine the individual's stance on politics or social issues).

Besides the problem of unlabeled data, our work in part relates to the 'scalability' and 'sparsity' challenges highlighted by prior research in recommender systems (Vozalis & Margaritis, 2003). Particularly, recommender systems should scale with the number of users and the number of available products. Toward this end, the proposed framework generates results from nearly 49,000 panelists and 600 programs in ~30 min on a desktop computer for runs with a few thousand genetic operations. Furthermore, the average user usually purchases or rates a limited number of items, which results in sparse user-item matrices. In the context of our work, too, a panelist is a fan of limited number of TV programs, which makes the panelist-show data sparse (we will discuss the data in detail in Section 3). The genetic algorithm framework presented here works well with sparse data as well as shown by the execution time and fitness values from the results. The fitness computations use methods from Python's NumPy implementation and are not customized for sparse data specifically.

3. Data

We were provided access to the Nielsen's national database on TV viewership for the entire US panel throughout 2016. The (de-identified) TV viewership data is at the household level; that is, the household daily viewership of all televised broadcasts (hereafter 'telecasts') from both broadcast and cable stations were tracked and recorded in minutes. There were nearly 1.5 million telecasts in 2016 and the households in Nielsen's national database throughout the US watched almost seven billion minutes of the telecasts.

We use program rankings provided by IMDB, TV.com, Nielsen, and TVGuide to compile the list of popular episodic programs nationwide in the US in 2016. The resulting list contains six hundred popular programs, all of which were broadcast to all states throughout 2016. Of note, the programs in the compiled list are not constrained to politics; they rather belong to a wide range of genres such as game, drama, talk, cooking, sports, news, and reality shows.

We draw on the share-of-wallet concept to define a feature that helps determine if a household can be considered as a fan of an episodic program. Specifically, we define "Average-Attention-Share" as the percentage of a household's (total) television viewership time that has gone to a specific episodic program. Toward this end, we first compute the total time each household spent watching television throughout 2016. We then compute the portion of the household's total watch time that went to the relevant telecasts of a specific episodic program. For example, if the Average-Attention-Share feature for < household A, *Planet Earth* > is 0.12%, it means that 0.12% of all the time that household A spent watching television throughout 2016 was devoted to the *Planet Earth*'s telecasts. A threshold on this feature further defines if the household is a fan of the episodic program.

Table 1 provides an example of the resulting data generated with an arbitrary threshold (e.g., 0.1%) on the Average-Attention-Share feature that we computed for every household-show from Nielsen's national database. Given this threshold, the first household in the data is considered a fan of *Basketball Wives* (i.e., more than 0.1% of the household's TV time was devoted to watching *Basketball Wives*, hence the corresponding flag is set to '1'), whereas the same household is not considered a fan of *Duck Dynasty*.

It is important to note that 'political preference' in Table 1 is not part of the (unlabeled) data; we only include it in the table as examples that help explain the problem statement and framework later in the paper. The resulting data (that will be used in the framework) contains 48,816 rows (one row per household), 600 columns (one column per program), and one column for the state (as the only demographic data). It should be noted that applying different thresholds on the computed household-show Average-Attention-Share feature generates different data. We will accordingly apply the framework to different data (generated with different thresholds) to determine (*i*) the set of programs with affinities to political and social issues, and (*ii*) the viewer political and social preferences.

The schema in Table 1 in part relates the present study to the data sparsity problem in recommender systems research (see Section 2.2). Specifically, the framework in this study estimates user political/social

Fig. 1. A chromosome (color online).

preferences—thereby building a recommender system that can, for example, suggest television advertisements that best match the viewer political attitude. However, given the large number of programs (and the threshold cut-offs), the data in Table 1 can be sparse, which is a significant problem in building recommender systems (Burke, 2002). The proposed genetic algorithm works well even in the presence of such sparse data and in part contributes to this common thread in recommender systems research (see, for example, Inan, Tekbacak, & Ozturk, 2018).

4. Formalism and problem statement

At the heart of our work is the problem of learning individual political and social preferences from unlabeled data, but in the presence of aggregate information. In this section, we use the US politics as an illustrative example to explain our unique formalism and problem statement

Let $P = \{P_1, P_2, \ldots, P_A\}$ be a set of *unobserved/unlabeled* individual preferences (e.g., $P = \{$'*Conservative*', '*Liberal*', '*Other*'$\}$). Also, let $S = \{S_1, S_2, \ldots, S_T\}$ be a set of episodic programs, and $H = \{H_1, H_2, \ldots, H_C\}$ be a set of households for whom the TV viewership data is available.

Further, let $S(H_k)$ be the set of episodic programs deemed as 'chosen' or 'watched' by H_k. To determine the watch (or choice) event, for now we use a threshold of 0.1%; that is, if a household's Average-Attention-Share (see Section 3) for an episodic program is greater than 0.1%, then that household is considered a fan of that program (we will conduct a sensitivity analysis on this cut-off).

Finally, let $P(H_k) \subset P$ be the (unlabeled) political preference of H_k. In Table 1 for instance, the last column represents the unobserved $P(H_k)$ values.

We seek to discover baskets of programs that correlate with P. In the context of the US politics, an illustrative example, liberal basket of shows (i.e., $B_{Democrat}$) might be {*Anderson Cooper*, *Basketball Wives*}, while a possible conservative basket (i.e., $B_{Republican}$) might be {*Sean Hannity*, *Duck Dynasty*}. Intuitively, households that watch more shows from $B_{Democrat}$ than $B_{Republican}$ might be interested more in Democrats (Liberals) than Republicans (Conservatives). We formalize this using an inferred preference rule to label the unlabeled data for household H_k based on which basket overlaps more with the household's choice set. More precisely, the preference rule assumes:

$$P(H_k) = \text{argmax}(S(H_k) \cap B_i) \tag{1}$$

In the data outlined in Table 1 (Section 3), this inferred preference rule would label the last household as '*Republican*' since that household watched more shows from the conservative basket than the liberal basket. The preference rule would label the first two households as '*Other*' given that the overlap with both the baskets is the same for the panelists.

How do we know if the discovered baskets are meaningful? Since we lack individual preference labels, we seek to maximize an objective function that considers how well the inferred panelist labels roll up to create *aggregate* statistics, which can be further compared to the actual aggregate vote shares (or other public opinions). This is basically the process through which the framework achieves one of the two goals we outlined in the introduction; that is, estimating a TV viewer's stance on politics or other public matters

In the US politics example, specifically, let $D\%(s)$ and $R\%(s)$ represent the actual percentage of votes that went to Democrats and Republicans respectively in state 's'. Given our access to the 2016 Nielsen data, in this paper we use the aggregate vote shares from the 2016 POTUS election. Based on the labeling of all households through the preference rule above, $ID\%(s)$ and $IR\%(s)$ represent the *inferred* percentages of votes that went to Democrats and Republicans in state 's' respectively.

Continuing the political example, we seek to learn baskets that minimize the error between the actual and inferred percentages of votes, averaged across all states. Specifically, we seek to discover sets $B_{Democrat}$ and $B_{Republican}$ such that $Avg(|(D\% - R\%) - (ID\% - IR\%)|)$ averaged across all states is minimized, and $(B_{Democrat}, B_{Republican} \subseteq S)$.

The above problem is NP-hard given the exponential search space (i.e., $\sim 2^{600}$ possible program subsets). In this paper we propose a framework based on genetic algorithms to illustrate how our problem formulation can be applied to developing practical methods to solve an important problem with broad implications in politics, social issues, and business.

5. The proposed genetic algorithm framework

Genetic algorithms (Goldberg & Holland, 1988) are a class of evolutionary algorithms broadly classified as *meta*-heuristics. Genetic algorithm have been used to generate heuristic solutions to NP-hard problems and have shown considerable value in applications characterized by complex search spaces (for an example in marketing, see Liu & Ong, 2008).

Genetic algorithms operate by first initializing, and then evolving, a population of "candidate solutions" to a problem. The "initialization" generates random candidate solutions of a certain number (i.e., "population size"). The "evolution" is based on a process that mimics natural selection, where "fitter" chromosomes are given higher probabilities to be selected for reproduction (i.e., crossovers and mutations) operations. These operations essentially take two candidate solutions and combine them to create two new offspring solutions. "Fitness" itself, needs formal specification and typically corresponds to the specific problem that the genetic algorithm is designed to solve.

In this paper, we define a chromosome as a string of *2*k* numbers, where the first '*k*' integers concern episodic programs (represented as numbers) that define $B_{Liberal}$ while the second '*k*' integers define $B_{Conservative}$. This representation fixes the basket size to a maximum of '*k*' episodic programs. This is primarily for expositional convenience and in practice does not constrain the solutions obtained since we run the GA with different values of '*k*' to learn baskets of potentially different sizes.

Fig. 1 provides an example of a chromosome in this representation where $k = 3$. The blue and red genes in the example in Fig. 1 represent Liberal (Democrat) and Conservative (Republican) programs in the chromosome respectively.

The "solution" implied by the chromosome in Fig. 1 is specifically:

- $B_{Democrat}$ = {*Anderson Cooper*, *Basketball Wives*, *The Daily Show*}
- $B_{Republican.}$ = {*The Sean Hannity Show*, *Duck Dynasty*, *The O'Reilly Factor*}

We define the "fitness function" f of a chromosome as:

$$f = 1 - Avg(|(D\% - R\%) - (ID\% - IR\%)|) \tag{2}$$

The intuition behind choosing the fitness function in (2) is that if a chromosome represents the "true baskets," then it should result in a labeling of households that closely corresponds to the actual (aggregate) voting percentages seen in each state. The fitness function follows the GA convention that higher fitness values are better.

The actual computation of the fitness function in the GA involves the followings:

```
Initialize population()

for evol = 1 to ITER:
# number of iterations of evolutionary operations
     If (random.uniform(0,1) < RANGEPROB:
               range_crossover()
     else:
               single_position_crossover()
     if random.uniform(0,1) < MUTATIONPROB:
               mutate()
     if (evol % 100 == 0)
               boost_diversity()
```

Fig. 2. Overview of the GA framework.

- Every individual is labeled as '*D*' or '*R*' based on how much their viewership overlaps with the $B_{Democrat}$ and $B_{Republican}$ sets implied by a chromosome. The label is based on the preference rule (Equation (1) in Section 4) that checks if the household watches more episodic programs in one basket than the other.
- From the labels, aggregate votes are computed for each state, from which the average inferred Democratic and Republican percentages (i.e., *ID*% and *IR*%) are computed.
- The aggregates are then compared to the actual vote percentages in each state to compute the fitness function (Equation (2)).

The evolutionary operators are defined in the standard manner. We consider two different crossover operations that swap "genes" either based on segments or specific positions. In *crossover-1*, particularly, we create two "children" chromosomes from two "parent" chromosomes by randomly selecting a gene and swapping its contents. In *crossover-2*, we pick a random point, but crossover entire segments, resulting in *child-1* having its first half of genes from *parent-1* and its second half of genes from *parent-2*. Genetic algorithms often try keeping these operations as simple and domain independent as possible—a norm that we follow here as well. Mutations are designed to simulate random "shocks" and can be implemented in different ways. In this paper, we implement this by picking a chromosome and randomly swapping a program with a different one.

Also as conventionally done, we simulate the natural selection process by using the fitness values of chromosomes as the likelihood of selecting them as parents in an evolutionary process. As a result, "fitter" chromosomes will have a greater chance of passing their genes to the next generation.

Finally, we implement a "diversity booster" within the genetic algorithm as a computational mechanism to replace several chromosomes in the population with newer ones that are generated randomly. This is usually done to create a more diverse gene pool after many iterations of evolutions to prevent premature convergence to a local optimum (Rocha & Neves, 1999). Fig. 2 presents the high-level GA framework.

6. Results and discussion

This section explains the results of applying the proposed framework to the 2016 television viewership data. Section 6.1 particularly presents the findings from the 2016 US presidential election results, while Section 6.2 explains the findings from the 2016 public opinions on controversial issues, namely global warming, same-sex marriage, and abortion. Section 6.3 discusses the findings in detail.

6.1. Television viewership and US politics

We first apply the framework to the 2016 television viewership data and US presidential election results to estimate the viewer political attitudes (i.e., Democratic or Republican) and discover the baskets of programs that are most watched by supporters of each party. We run the GA for different (maximum) basket sizes, from two to six programs per

Table 2
Baskets of television programs and political affinity (the best basket and fitness value are shown across different settings (rows) of the GA).

Max Size	Cut-off	Best chromosome		Fitness
		Democratic genes	Republican genes	
2	0.01	Modern Family | Pioneer Woman	College Football | Reba	0.927
	0.05	Jimmy Kimmel | Diners Drive-Ins & Dives	College Football | 2 Broke Girls	0.920
	0.1	Anderson Cooper | Modern Family	College Football | FNC The Five	0.922
	0.2	Modern Family | Sports Center	College Football | 48 Hours	0.915
	0.3	Anderson Cooper | Saturday Night Live	College Football | American Pickers	0.917
3	0.01	Modern Family | Shades of Blue | MSNBC Hardball	College Football | FNC Fox & Friends | Bonanza	0.932
	0.05	Chicago Fire | SportsCenter | Yes to the Dress	College Football | Reba | The Young & the Restless	0.933
	0.1	Anderson Cooper | Modern Family | America's Got Talent	College Football | FNC The Five | Price is Right	0.928
	0.2	Modern Family | Saturday Night Live | Kardashians	College Football | Price is Right | First 48	0.927
	0.3	Anderson Cooper | Saturday Night Live | Chopped	College Football | American Pickers | 48 Hours	0.921
4	0.01	CNN Jake Tapper | Chicago PD | How I Met Your Mother | Deadliest Catch	College Football | Duck Dynasty | Pardon the Interruption | Charmed	0.927
	0.05	Anderson Cooper | Modern Family | 60 Minutes | My 600-lb Life	College Football | FNC Fox & Friends | Criminal Minds | Crime Time	0.932
	0.1	Anderson Cooper | Dancing with the Stars | Cops| The Blacklist	College Football | FNC The Five | Price is Right | FNC For the Record w/ Greta	0.928
	0.2	Anderson Cooper | Modern Family | Chopped | NFL Sunday	College Football | Price is Right | 48 Hours | Fixer Upper	0.924
	0.3	Good Morning America | SportsCenter | 60 Minutes | Chopped	College Football | American Pickers | Forensic Files | Criminal Minds	0.917
5	0.01	MSNBC Rachel Maddow | CNN Jake Tapper | Law and Order | Fresh off the Boat | Parts Unknown	Duck Dynasty | Bonanza | Chrisley Knows Best | Big Brother	0.932
	0.05	Blindspot | Jimmy Kimmel | Chicago PD | Chopped | Catfish	College Football | Reba | Leverage | First 48 | Code Black	0.932
	0.1	Anderson Cooper | Grey's Anatomy | Walking Dead | Shark Tank | Diners Drive-Ins and Dives	College Football | FNC Fox & Friends | Forensic Files | The Middle | Let's Make a Deal	0.933
	0.2	Anderson Cooper | Dancing with the Stars | Kardashians | NBC Nightly News | Good Morning America	College Football | Price is Right | Fixer Upper | Cops | FNC The Kelly File	0.923
	0.3	Anderson Cooper | Today Show | NBC Nightly News | 60 Minutes | SportsCenter	College Football | The Voice | Forensic Files | NBC Nightly News | NCIS	0.915
6	0.01	MSNBC Hardball | Yes to the Dress | Designated Survivor | Star Trek | CNN Fareed Zakaria | Chicago PD	College Gameday | Bonanza | FNC Outnumbered | Limitless | Lethal | Weapon | Fast 'n Loud	0.932
	0.05	American Idol | Chopped | Dancing with Stars | Chicago Fire | CNN Jake	College Football | FNC Outnumbered | Let's Make a Deal |	0.933

(*continued on next page*)

Table 2 (*continued*)

Max Size	Cut-off	Best chromosome: Democratic genes	Best chromosome: Republican genes	Fitness
		Tapper | America's Got Talent	Supernatural | The Middle | Survivor	
	0.1	Modern Family | Anderson Cooper | Madam Secretary | The Real Housewives | Dancing with the Stars | Diners Drive-Ins and Dives	College Football | FNC The Kelly File | Pawn Stars | Last Man Standing | The Talk | The Young & the Restless	0.937
	0.2	Anderson Cooper | Dancing with the Stars | Kardashians | 20/20 | NBA | NFL Sunday	College Football | Price is Right | 48 Hours | Forensic Files | Criminal Minds | The Voice	0.927
	0.3	Anderson Cooper | Saturday Night Live | Property Brothers | NBA | 60 Minutes	College Football | CSI | 48 Hours | Fixer Upper | Criminal Minds	0.917

basket (see Section 5). In each run, the population size and the number of iterations are set to 300 and 5,000 respectively.

Table 2 presents a summary of the discoveries, showing the best baskets discovered under different settings; i.e., across different basket sizes along with a sensitivity analysis on cut-off thresholds (i.e., {0.01, 0.05, 0.1, 0.2, 0.3}) that are used to consider a household as a fan of an episodic program. In other words, we run the GA for different basket sizes on different data generated by a range of thresholds applied to the household-show Average-Attention-Share data (see Section 3).

On average, the discovered baskets result in an average prediction within 7% of the margin of the actual election results, which has to be viewed in the wake of a limitation: in most of the states, less than one thousand panelists serve as proxies for an entire state's population. Nonetheless, the directional predictions were correct in approximately 85% of the states. Also, a random prediction (between 0 and 100%) of the margin would have been off by over 60%, while predicting the average margin (-3.67%, based on the average vote margin across all states in the data) would also have been off by 18%. Given these baselines, the average fitness of 0.93 achieved by the framework suggests that the discovered baskets are capturing potentially strong signals related to individual viewer political attitudes.

Except *Saturday Night Live*, and many CNN, MSNBC, and Fox News Channel (FNC) programs, which clearly advocate left and right-wing politics and are correctly assigned to the relevant Democrat and Republican baskets by the algorithm, the shows in Table 2 might seem unrelated to politics at face value. However, several of the seemingly unrelated programs discovered by the framework also appear in the lists that Experian (2010, 2012) compiled based on the viewers political party registrations in their National Consumer Study, and the findings in the Norman Lear Center study (Blakley et al., 2019). Below are some of the episodic programs that appeared frequently in the discovered baskets with their studied connection to the US politics.

- *College Football* has the highest concentration of "ultra Conservative" fans among all non-news programs (Experian, 2012). Analysis of Google Trends data (Paine, Enten, & Jones-Rooy, 2017) also indicates that among seven major sports, *College Football* search traffic is most positively correlated with Trump's vote share in the US media markets (also, Leonhardt, 2014). The program appears in almost all Republican baskets discovered by the framework.
- *The Price is Right* is a game show where contestants compete to win prizes by guessing the pricing of merchandise. The program, however, has been among the top 10 shows with the highest percentage of Republican advertisements spots (Scott Diamond, 2014). The show is the second most frequent program in the discovered Republican baskets by the framework.
- *Pawn Stars* is an American reality series about a family-run pawnshop. The show appears in the list of top eight TV shows that "Conservatives will love" (Hawkins, 2018) and is one of the most-watched shows by Conservative viewers (Blakley et al., 2019). The program is also discovered by the framework.
- *Modern Family* is the ABC's progressive comedy that features a gay married couple and their adopted daughter. *Modern Family* is a primary program in reaching "super Democrats" (Experian, 2012) and is the show that Liberal viewers enjoy watching (Blakley et al., 2019). The show is in several Democratic baskets discovered by the framework.
- *The Middle* is an American sitcom about the daily mishaps of a middle-class, middle-America family. In an annual research survey that measures the consumer preferences of various political ideologies (Hibberd, 2011), the program appears among more traditional comedies with appeal for conservative viewers.

Section 6.3 will further discuss these and other findings and their implications in detail.

Fig. 3. Fitness evolution paths.

Table 3
State-level opinion statistics.

Social issue	One ('Pro') side	The other ('Anti') side
State-level opinion on climate change by Yale Program on Climate Change Communication (Howe, Mildenberger, Marlon, & Leiserowitz, 2015).	Estimated percentage who think that global warming is happening.	Estimated percentage who do not think that global warming is happening.
State-level opinion on same-sex marriage by The American Values Atlas (Cox et al., 2017) (; Public Religion Research Institute (n.d.)., 2019	Estimated percentage who strongly favor allowing gay and lesbian couples to marry legally.	Estimated percentage who strongly oppose allowing gay and lesbian couples to marry legally.
State-level opinion on abortion by Pew Research Center (n.d.)., 2019	Estimated percentage who believe abortion should be legal.	Estimated percentage who believe abortion should be illegal.

Table 4
Pearson correlation coefficients between state-level statistics.

	Global warming happening	Same-sex marriage	Abortion rights
State-level Democrats vote margin	0.95	0.77	0.68

6.1.1. Diversity boosters and impact on evolution

As common in genetic algorithms, we see the average fitness of the population converging soon to close to the fitness of the best chromosome. We accordingly modify the genetic algorithm to focus on diversity more during the process by randomly regenerating subsets of chromosomes. As expected, the fitness evolution paths (Fig. 3) show that the population on average might be seeing newer and more random genes, which illustrates several powerful levers that the genetic algorithm framework can bring to the discovery process.

6.2. Television viewership and US social issues

In addition to the 2016 presidential election, we apply the genetic algorithm framework to the public opinion statistics on social issues, namely global warming, same-sex marriage, and abortion (Table 3 outlines the statistics). The methodology is similar to the one we practiced in the context of presidential elections. Again, the panelist (individual) attitudes to these social issues are unlabeled, and the only available data concerns TV viewership and state-level public opinion statistics.

The social issues explored in this study also have connections to politics in the US. Historically, political conservatism in the US has been negatively correlated with support for same-sex marriage and abortion rights (for a review, see Sherkat, Powell-Williams, Maddox, & De Vries, 2011). Recently, global warming has also become a political topic in the US, with some prominent conservatives rejecting the scientific consensus (among 97% of climate scientists; see NASA, 2019) on climate change. These correlations also resonate in our data, particularly between the 2016 state-level Democrats' vote margin and the opinion margins on the social issues. Table 4 contains the corresponding Pearson correlation coefficients, all of which are significant at $\alpha = 0.0001$. Given this correlation, it might be interesting to study the discovered baskets for possible overlaps and differences with the ones learned previously where the focus was specifically politics.

Toward this end, we apply the framework to the viewership data that are generated by different thresholds (which defines if a household is a fan of a program; see Section 3), and public opinion statistics on global warming, same-sex marriage, and abortion rights. For each social issue, specifically, we run the algorithm for different maximum basket sizes, from two to six programs per basket. We also conduct a sensitivity analysis on the viewership thresholds (i.e., {0.01, 0.05, 0.1, 0.2, 0.3}). The population size and the number of iterations are again set to 300 and 5,000 respectively in each run. In total, therefore, the framework discovers 25 'Pro' and 25 'Anti' baskets for each social issue.

Algorithmically the approach here is identical to what was done earlier in the paper. Each anonymous panelist is labeled as 'Pro' if their viewership overlaps more with the shows in the 'Pro' basket than the 'Anti' basket. Based on this labeling, aggregate statistics from this panel can be compared to the actual 'Pro' opinion percentages across states in order to determine the fitness of a basket.

We present the detailed findings for each social issue in Online Appendix A. For ease of exposition, however, we compute the number of times each program is contained within the Pro and Anti baskets with regard to each social issue. The plots in Fig. 4 use the computed frequencies to contrast the top ten most frequent programs in these baskets. The left (blue) programs in the plots are contained in 'Pro', and the right (red) ones are contained in the 'Anti' baskets.

To illustrate, Fig. 4.a indicates that the most frequent program in the Democratic baskets for the 2016 presidential election is *Anderson Cooper* on CNN, which stands against *College Football* as the most frequent program in the Republican baskets (*Anderson Cooper* is contained in 52% of Democratic baskets, and *College Football* is contained in 92% of Republican baskets). The second episodic program in the list of Democratic shows is *Modern Family*, which contrasts *The Price Is Right* in the list of Republican shows. With regard to global warming, Fig. 4.b indicates that the most frequent show in the "Pro" baskets is *Law and Order*, which stands against *The O'Reilly Factor* on Fox News Channel.

Given the correlations between the social issues and conservatism/liberalism, we expected the framework to find similar programs to the ones in politics, yet the algorithm did find several programs that were not contained in the corresponding baskets for the presidential election. Table 5 indicates, for example, that *NASCAR* is specifically a frequent show in the 'Anti' baskets for public opinion on global warming (*NASCAR* has been criticized for its carbon footprint), although it is not part of any Republican basket for the presidential election. In a similar vein, *Fast N' Loud* (in which the reality show crew search for tired and run-down cars to restore them for profit) and *Street Outlaws* (which provides an inside look into the American street racing) are also car shows that are specifically contained in the 'Anti' baskets for climate change (and not in the Republican baskets). Furthermore, *Hannity* and *The O'Reilly Factor* are programs on Fox News Channel that appear frequently in the 'Anti' baskets for climate change, which is interesting given some of the conservatives' rejection of the scientific consensus on global warming.

Finally, in the survey conducted by Blakley et al. (2019), *Law and Order* and *The Big Bang Theory* are both among the shows that Liberals watch more than other viewers. As the plots in Fig. 4 indicate, *Law and Order* appears in 92% of the 'Pro' baskets that the framework discovered for global warming. In a similar vein, the framework assigned *The Big Bang Theory* to 20% of the 'Pro' baskets for same-sex marriage.

6.3. Discussion

The alignment between the discovered baskets and the programs with appeal for registered voters (from both parties) notwithstanding, there are a few shows that are deemed to have appeal for Republican viewers and yet are assigned to Democratic baskets by the framework (Table 2). *Dancing with the Stars*, for example, is considered a Republicans' favorite show by Experian (2012); however, the framework assigns the program to a few Democratic baskets. *Diners, Drive-Ins and Dives* is another program that was suggested to have appeal for Republican viewers and yet is assigned to a few Democratic baskets by the framework.

While the correct assignment of several Conservative and Liberal programs to the corresponding baskets in part corroborate the findings, the above differences (between the discovered baskets and the marketing research groups' lists) both highlight a limitation, and put stress on

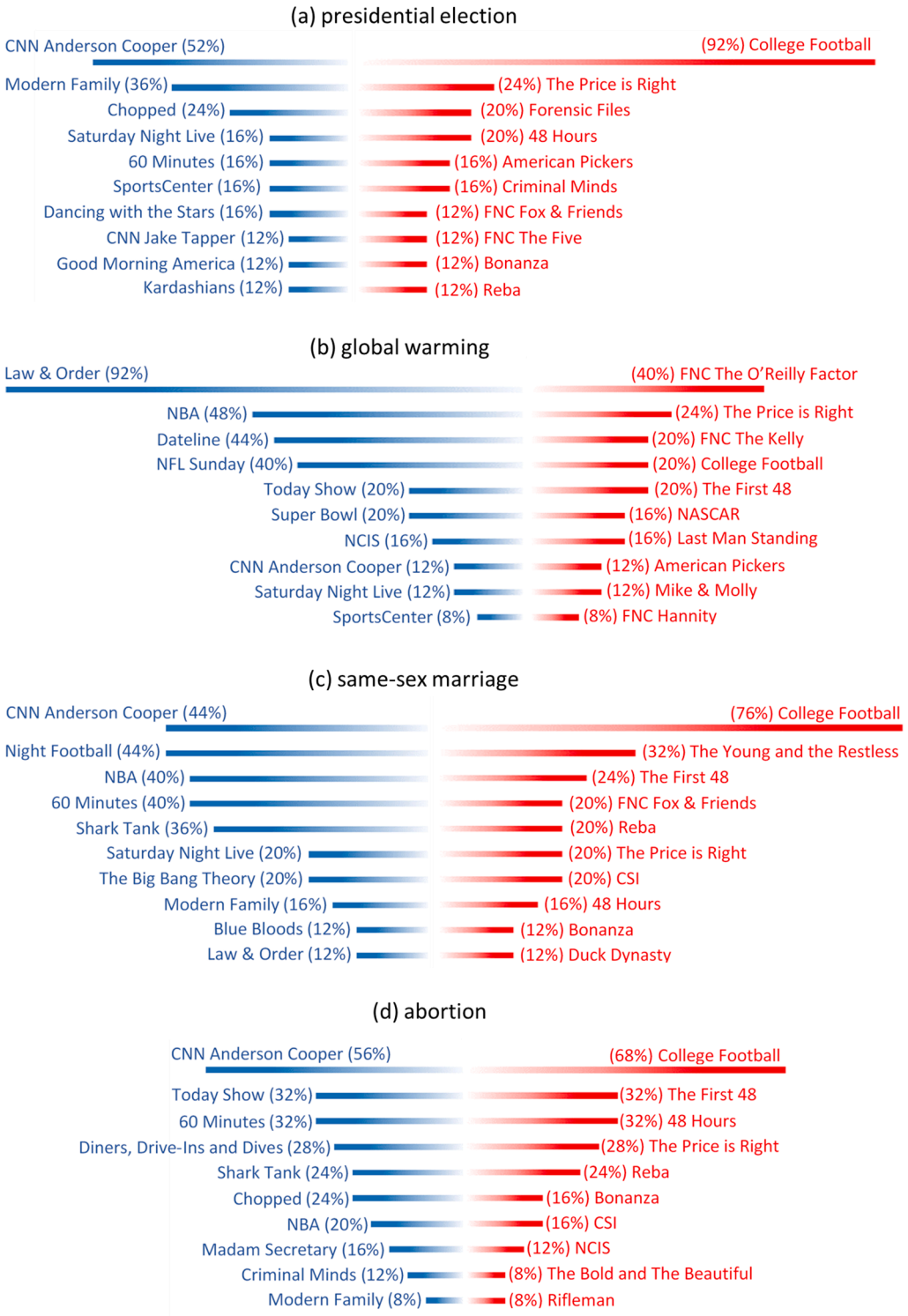

Fig. 4. (color online). Top ten episodic programs across baskets

the proper interpretation of the algorithm findings. As a limitation, marketing research groups (e.g., Experian) usually conduct their "cross-sectional" studies using limited samples, while program themes and ideological makeup of their audience can shift in the long run. For example, Experian (2012) found *Dancing with the Stars* to be a Republicans' favorite, while Blakley et al. (2019) findings suggest that the program is the purple (swing) viewers most-watched show. In the case of *Diners, Drive-Ins and Dives* (which Experian suggested to have appeal for Republicans and yet was assigned to a few Democratic baskets by our framework), the show star had mocked a campaign slogan (i.e., "#MakeFlavortownGreatAgain") in 2020 to announce his nomination for the upcoming election.

To interpret the findings, the discovered shows should be viewed as part of (larger) baskets of programs that the framework finds to have strong signals for the individual viewer political attitudes—which again puts stress on the fact that the algorithm is estimating the individual preferences from aggregate data. To illustrate, the findings suggest that a viewer who at the same time is a fan of *Anderson Cooper*, *Kardashians*, *20/20*, *Dancing with the Stars*, *NBA*, and *NFL Sunday* is more likely to have Democratic attitudes (see Table 2, basket size = 6, threshold =

Table 5
Frequent shows specific to global warming, same-sex marriage, and abortion.

	"Pro-Issue" genes	"Anti-Issue" genes
Climate change	Dateline, Super Bowl, NCIS, Night Football, The Bachelor	NASCAR, The O'Reilly Factor, Hannity, Gunsmoke, Mike & Molly, America's Got Talent, Tonight Show
Same-sex marriage	The Big Bang Theory, Blue Bloods, What Would You Do?, Night Football	Days of Our Lives, American Ninja Warrior, CSI, Dr. Phil
Abortion rights	The Good Place, Criminal Minds	Rifleman, The Bold and The Beautiful, CSI, NASCAR

0.2). Here, *Dancing with the Stars* (which is suggested to have appeal for Republicans by Experian, 2012) is only one show in this basket that has six programs of different genres.

To discuss this further, we design an aggregate-level measure that indicates the popularity of an individual program in every state. For each program, particularly, we compute 'Minutes Per Household', which captures the time that an average household in a state spent on watching a specific program throughout 2016. The resulting high-dimensional data has 600 columns (one per program) for 51 states (including District of Columbia). This popularity measure allows us to examine the correlation between the aggregate viewership of every program and the vote share.

Using the new aggregate-level popularity measure, we first find the top five popular programs in each state. From this, we subsequently count the number of times that a program appears in the lists of top five popular programs in the states that voted for Democrats or Republicans in 2016. The plots in Fig. 5 suggest that, for example, *Law and Order* was simultaneously among the top five popular programs in all 30 states that voted for Trump and all 21 states that voted for Clinton. The plots indicate a significant overlap between the most popular shows in the Democratic and Republican states in 2016. Interestingly, *Fox and Friends* appears twice among the popular programs in the Democratic states, while the program does not belong to any popular list in the Republican states (the program, however, was correctly assigned to the Republican baskets by the framework; see Table 2).

In a similar vein, we compute the Pearson coefficient of correlation between every program's Minutes Per Household and the Democrats' and Republicans' vote margins in 2016. Table 6 presents the programs whose (aggregate-level) popularity measure is highly correlated with the parties' vote margins, along with their Pearson correlation coefficients (all of which are significant at $\alpha = 0.0001$). Again, while there are some obvious shows (e.g., Fox News Channel programs) in Table 6, the findings are not particularly informative about the viewership of combinations (sets/baskets) of programs and how such viewership determines individual political or social preferences. This is in part because the findings in Table 6 are coming from separate analyses of the shows at the aggregate-level (e.g., correlation between the aggregate-level popularity of an individual show—in isolation from the remaining 599 shows—and the vote margin). A corollary of the findings in Table 6, for example, is that if a viewer watches all Fox News Channel shows along with *Jep and Jessica* (which is a *Duck Dynasty* spin-off), that individual is most likely leaning towards the Republican party, which is obvious (yet highly unlikely that a viewer is only a fan of such combination).

7. Incorporating Multi-Objective settings

The framework we presented shows how a genetic algorithm that seeks to optimize a single fitness function (e.g., on vote margin) can be used to learn baskets, and in turn individual preferences, from aggregate data. For the specific context considered (e.g. individual political preferences), it is also possible to frame this in a multi-objective setting. This section considers this scenario and presents three approaches based on ideas in the Multi-Objective Evolutionary Algorithms (MOEA) literature (Konak et al., 2006; Emmerich & Deutz, 2018). Specifically, we define two objective functions in a multi-objective setting:

$$f_1 = 1 - Avg(|D\% - ID\%|) \tag{3}$$

$$f_2 = 1 - Avg(|R\% - IR\%|) \tag{4}$$

Essentially, each of the two objectives aims at getting the baskets defined such that, on average, the inferred Democratic (Republican) percentage is as close as possible to the actual Democratic (Republican) percentage in the states.

The literature in the MOEA area presents different approaches that can be used in such settings (Konak et al., 2006; Emmerich & Deutz, 2018). At the heart of these ideas is the notion of using evolutionary computing to learn a pareto frontier, which contains the set of solutions that are non-dominated by other solutions. To this end, the population of chromosomes represents a set of solutions that is expected to approximate this pareto frontier.

To demonstrate how these ideas can extend the proposed framework in a

Table 6
Most correlated programs with politics at aggregate-level.

TV programs with viewership highly correlated to:	
Democrats' vote margin	Republicans' vote margin
The Daily Show (0.74)	Jep & Jessica (0.72)
Real Time (0.66)	Moonshiners (0.63)
Last Week Tonight (0.62)	FNC Hannity (0.63)
Ballers (0.60)	FNC The Kelly File (0.61)
The Night Of (0.60)	FNC On The Record (0.61)
The Nightly Show (0.59)	The O'Reilly Factor (0.60)
Veep (0.57)	Last Man Standing (0.59)
Vinyl (0.56)	Fast N' Loud (0.59)
Broad City (0.55)	Diesel Brothers (0.59)
Insecure (0.55)	Alaskan Bush People (0.58)

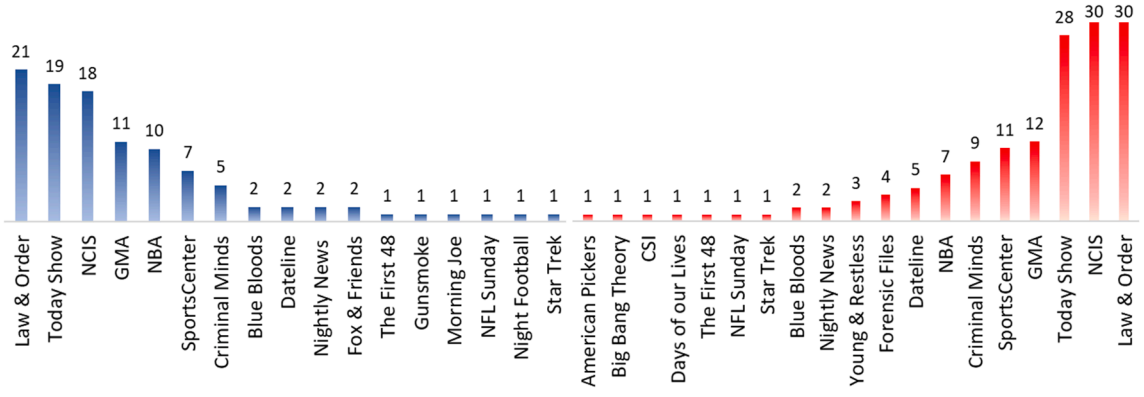

Fig. 5. (color online). Most popular programs in 2016.

Table 7
Solutions in the Pareto Frontier.

MOEA Method	Chromosomes on the Pareto Frontier		Fitness	
	Democratic genes	Republican genes	Dem	Rep
Scalarization	The View | Law & Order Good Morning America | Super Bowl	Criminal Minds | Forensic Files |The First 48 | American Pickers	0.95	0.78
	Bones | Modern Family Family Guy | CNN Anderson Cooper	Voice | CSI |Law & Order | Naked and Afraid	0.74	0.95
	Diners, Drive-Ins & Dives | Kardashians | NFL Sunday | Super Bowl	Bones | NCIS |Castle	0.92	0.90
RWGA	Chopped | Super Bowl Law & Order | The Walking Dead	The First 48 | American Pickers | 48 Hours	0.95	0.82
	CNN Anderson Cooper | Dancing with the stars | CNN Erin Burnett | Saturday Night Live	Last Man Standing | Bones |Big Bang Theory | Law & Order	0.80	0.94
	CNN Erin Burnett | Impractical Jokers | NFL Sunday | Shark Tank	American Funny Videos | 20/20 |48 Hours	0.89	0.88
VEGA	20/20 | Property Brothers | CNN Anderson Cooper | Impractical Jokers	FNC O'Reilly Factor | FNC Kelly File | Fox & Friends | NASCAR	0.94	0.72
	CNN Erin Burnett | Jimmy Kimmel | CNN Anderson Cooper | 60 Minutes	College Football | Undercover Boss | Fox & Friends | Law & Order	0.76	0.96
	CNN Anderson Cooper | Dateline | Diners, Drive-Ins & Dives | Bones	College Football | Bachelor | Fixer Upper | Blue Bloods	0.86	0.84

multi-objective context we implemented three methods for MOEA: (*i*) Linear Scalarization, (*ii*) Random Weighted Genetic Algorithms (RWGA), and (*iii*) Vector Evaluated Genetic Algorithms (VEGA). Linear Scalarization, specifically, converts a multi-objective problem into a single objective by using weights. RWGA, on the other hand, maintains a population of solutions, where each solution's fitness is determined by a specific combination of (random) weights used for each object (in that sense, the population in RWGA simulates fitness across various combinations of the two objectives). Finally, VEGA maintains multiple sub-populations, where each population uses fitness defined based on one specific objective (and there are therefore as many sub-populations as objectives).

There are two important considerations here. First, how the "selection" process works to identify chromosomes that are used in the crossovers or mutations. Second, how the pareto frontier is generated/maintained. The multi-objective component in these methods is primarily tied to how selection (for evolutionary operations) is implemented. In RWGA, for example, the initial population is a random set; that is, solutions are selected for crossover/mutations based on a fitness function where each chromosome's weights for the different objectives are generated randomly. In VEGA, solutions are randomly sorted, and half the solutions use f_1 while the other half uses f_2 to define fitness. Chromosomes are selected for crossover and mutation in the usual manner (proportional to fitness); however, given the different sub-populations it is possible to choose parents who are essentially optimized for different objectives. Linear scalarization essentially mimics a single objective GA, but the weights are user-specified (and varied to generate different solutions).

In all these methods, once a certain number of offsprings are generated, the pareto frontier set is periodically updated. This update essentially takes the current pareto frontier and the new population of chromosomes and subsequently determines the new pareto frontier.

Table 7 presents three examples of the non-dominated solutions in the pareto frontiers generated by (*i*) Linear Scalarization with equal weight, (*ii*) RWGA and (*iii*) VEGA based on identical populations (N = 100), 2000 iterations of genetic operations, and threshold cut-off set to 0.1 (Section 3). Among the three examples presented for each method, the first two are the points at the extremes of the pareto frontier (highest fitness in one of the two objectives), and the third one is in the middle of the corresponding frontier. The findings presented in Table 7 indicate that these methods do have an advantage of being able to generate pareto frontiers (e.g., the best fitness values for the extreme points are between 0.94 and 0.96).

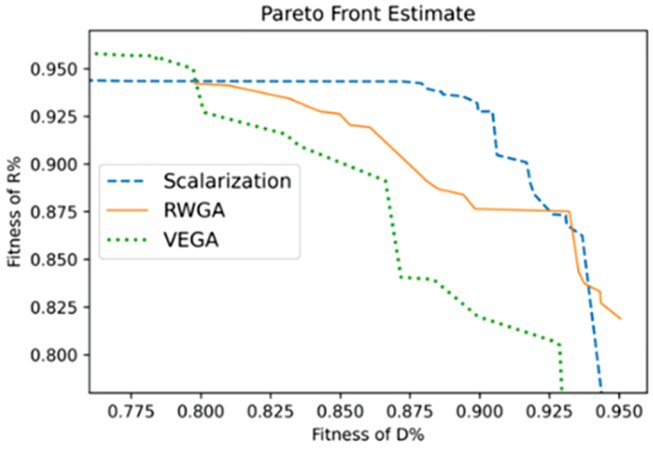

Fig. 6. Pareto Frontier Estimate from Scalarization, RWGA, and VEGA.

Fig. 6 presents the pareto frontiers generated by each of these methods. Each method generated approximately 25 solutions in the pareto frontier, and the plot presents the final set for each method. In this setting, the solutions in the scalarization frontier appear to dominate the others. However, both VEGA and RWGA show improvements at the extremes of the pareto frontier.

More generally, these findings point to the potential value of multi-objective evolutionary computation over the traditional GA. Rather than focus on generating a single best solution, the flexibility of having a pareto frontier to choose solutions from is appealing. However, the choice will be driven by a deep understanding of the problem being solved. In the application context of this paper, for example, the margin of victory in each state is a significant single objective, which is why this paper focused on the traditional GA. Yet, when using GAs to solve the broader problem of inferring individual preferences from aggregate data, having the ability to formulate this in a multi-objective setting as shown in this section is valuable.

8. Conclusions and limitations

In this paper, we present a genetic algorithm framework that uses unlabeled television watch data to learn baskets of programs whose viewership can determine the individual viewer political or social attitudes. We applied the framework to highly-granular TV viewership data of approximately 49,000 US households throughout 2016 along with the corresponding vote share and public opinion statistics (on global warming, same-sex marriage, and abortion). Throughout the process, the framework estimated individual viewer attitudes toward US politics and other public matters, the outcome of which was aligned with the 2016 US election results. Previous survey studies (e.g., Blakley et al., 2019; Experian, 2012) on television viewership and US politics in part corroborate the findings of the proposed framework in this study.

Our main contribution is a flexible general-purpose framework that can leverage *aggregate* data (e.g., vote shares or public opinions) and disaggregate data on a different variable (e.g., program viewership) to learn optimal sets of observable individual choice signals (e.g., baskets of shows) that predict microlevel unobservable/unlabeled preferences of interest (e.g., *individual* political and social preferences). In addition to advertising micro-targeting, the potential applications of this algorithm can be significant given the wide interest in using big data for

political and social campaigning and building election forecasting models with non-polling data. It should however be stressed again that our framework is minimally intrusive and only uses completely de-identified household TV viewing history. Specifically, the framework requires no demographic data except the household state. The only political and social data used by the framework are aggregate vote and public opinions data. While the setting considered in this paper is binary, we note that the chromosome representation can be extended to incorporate baskets corresponding to multiple classes as well (with no additional change required in our GA). This important aspect makes the proposed framework relatively general.

Methodologically, previous marketing research (Chen & Yang (2007); Musalem, Bradlow, and Raju (2008); (2009;))) has applied Bayesian analyses of aggregate data to augmenting individual brand choices from their posterior distribution. Our framework, too, estimates individual political/social preferences from aggregate statistics. However, in addition to individual preferences, our framework also learns another variable of interest (i.e., baskets of programs whose viewership can determine the individual's stance on politics or social issues). Furthermore, the suggested approaches in marketing research worked on a significantly smaller scale than what we have tested here using data that represents 48,816 panelists and 600 shows. We believe methods such as these offer researchers new alternatives to existing techniques and can help bring in novel computational approaches to the literature.

More generally, we view this as a broad framework that can be used in different contexts (other than politics) as well. For example, instead of the individual preference being political affiliation, one might consider scenarios where this could be a choice of car, or preference for golf. In such cases, all we need is aggregate-level data on such preferences (e.g., percentage of people who bought an electric car in each state). The discovered baskets may then be used to advertise products in programs that may be more likely to be watched by those who may buy electric cars. These are interesting directions for future work.

There are limitations to our work as well. First, the panel data we use is household-level viewership data. We do not have the ability to break this down into specific individuals in the household and hence treat each household as one entity. This adds noise to the data (and discoveries) since it is possible that the members of the households might have different political or social interests that show up differently as each of them watch their programs of interest. While we do not have the individual data, the vendors who track them certainly do, and can use the methodology in a more granular manner. Second, the panel data at our disposal is limited to one year; hence we only apply the proposed framework to a single presidential election. While we mitigate this limitation by applying the algorithm to public opinions on climate change, same-sex marriage, and abortion rights in the same election year, future work can examine the genetic algorithm in different years. This future research direction is particularly interesting as program themes and ideological makeup of their audience can shift in the long run. Third, while we showed how closely the aggregate voter statistics match actual voting behavior in the states, a large part of our exposition was appealing to other studied connections between the discovered shows and the outcomes of interest, making our findings and arguments more exploratory rather than confirmatory. Subsequent experiments targeting these baskets may be needed to provide confirmatory evidence on the value of the discovered signals. Fourth, given the large scale of our data did not implement Bayesian methods for comparison, but look to future work to tease out the relative merits of theory-driven Bayesian approaches versus computational heuristics from AI. Fifth, AI-based heuristics such as this have extensive levers and present many opportunities for hyper-parameter optimization as well. For ease of exposition, in this paper we made choices for parameter values and did not consider hyper-parameter optimization, which could possibly improve these results. The limitations notwithstanding, we believe the ideas presented here offer a compelling narrative and illustrate the potential of AI-based approaches for discovery in data.

CRediT authorship contribution statement

Balaji Padmanabhan: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Investigation, Supervision, Validation. **Arash Barfar:** Conceptualization, Data curation, Visualization, Formal analysis, Methodology, Writing - original draft, Investigation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors extend their gratitude to the editors and five anonymous reviewers for their detailed reviews and insightful suggestions.

Appendix A:. Detailed Listing of discovered baskets for social issues

.

Table A1
Pro and Anti baskets for climate change.

Max Size	Cut-off	Best chromosome		Fitness
		Pro-Issue genes	Anti-Issue genes	
2	0.01	Law & Order | Super Bowl	Gunsmoke | Mike & Molly	0.9688
	0.05	Law & Order | NBA	Fast N' Loud | Reba	0.9653
	0.1	Law & Order | Dateline	FNC Fox & Friends | Let's Make a Deal	0.9504
	0.2	Law & Order | Night Football	Hannity | Hannity	0.9213
	0.3	Law & Order | NFL Sunday	FNC The Kelly File | FNC The Kelly File	0.8954
3	0.01	CNN Anderson Cooper | Law & Order | NBA	Duck Dynasty | Last Man Standing | Two and a Half Men	0.9726
	0.05	Law & Order | Dateline | Super Bowl	Reba | The Talk | First 48	0.9630
	0.1	Law & Order | SportsCenter | Super Bowl	NASCAR | FNC The Five | The Young and the Restless	0.9581
	0.2	Law & Order | Dateline | NFL Sunday	The O'Reilly Factor | The Price Is Right | The First 48	0.9430
	0.3	Law & Order | Today Show | NBA	The O'Reilly Factor | The O'Reilly Factor | The O'Reilly Factor	0.9313
4	0.01	Modern Family | Elementary | NBA | Chopped	Gunsmoke | Big Brother | Dr. Ken | Teen Mom	0.9692
	0.05	CNN Anderson Cooper | Law & Order | NCIS | Shark Tank	Reba | Mike & Molly | American Idol | Scorpion	0.9667
	0.1	Law & Order | Saturday Night Live | NCIS | Super Bowl	College Football | Pawn Stars | 2 Broke Girls | The Young and the Restless	0.9615

(*continued on next page*)

Table A1 (*continued*)

Max Size	Cut-off	Best chromosome Pro-Issue genes	Best chromosome Anti-Issue genes	Fitness
	0.2	Law & Order | Today Show | Dateline | NFL Sunday	American Pickers | The Price Is Right | The First 48 | Tonight Show	0.9485
	0.3	Law & Order | Dateline | NBA | NFL Sunday	The O'Reilly Factor | FNC The Kelly File | CSI | The O'Reilly Factor	0.9353
5	0.01	CNN Anderson Cooper | Dateline | Bachelor | NBA | Cops	Duck Dynasty | Pit Bulls & Parolees | Brooklyn Nine-Nine | AFV | Crime Time	0.9711
	0.05	CNN Jake Tapper | Law & Order | 60 Minutes | NBA | Diners, Drive-Ins and Dives	FNC Fox & Friends | Impractical Jokers | Code Black | Castle | Survivor	0.9704
	0.1	Law & Order | Kardashians | NCIS | NBA | NFL Sunday	NASCAR | FNC For the Record w/ Greta | Pawn Stars | Last Man Standing | Colbert	0.9586
	0.2	Law & Order | Saturday Night Live | NFL Sunday | Dateline | NBA	College Football | The Walking Dead | The Price Is Right | Tonight Show | The First 48	0.9510
	0.3	Law & Order | Dateline | Today Show | NBA | NFL Sunday	The O'Reilly Factor | FNC The Kelly File | American Pickers | Forensic Files | The O'Reilly Factor	0.9434
6	0.01	Law & Order | Bachelor | Little Big Shots | Forensic Files | Cake Boss | NFL Sunday	The Bill Cunningham Show | Street Outlaws | The Curse of Oak Island | Gold Rush | Mike & Molly | Last Man Standing	0.9681
	0.05	The Situation Room | Saturday Night Live | Dateline | America's Got Talent | NBA | Property Brothers	NASCAR | Fast N' Loud | College Football | Survivor | Days of Our Lives | Star Trek	0.9715
	0.1	Law & Order | Good Morning America | Super Bowl | Let's Make a Deal | 20/20 | Today Show	NASCAR | The Price Is Right | 2 Broke Girls | The First 48 | The View | Last Man Standing	0.9644
	0.2	Law & Order | NFL Sunday | Law & Order | NCIS | NBA | Dateline	The O'Reilly Factor | College Football | The Price Is Right | The First 48 | America's Got Talent | The Price Is Right	0.9584
	0.3	Law & Order | Dateline | Today Show | SportsCenter | NFL Sunday | Night Football	The O'Reilly Factor | FNC The Kelly File | College Football | American Pickers | Forensic Files | America's Got Talent	0.9348

Table A2
Pro and Anti baskets for same-sex marriage.

Max Size	Cut-off	Best chromosome Pro-Issue genes	Best chromosome Anti-Issue genes	Fitness
2	0.01	MSNBC Rachel Maddow | Shark Tank	Bonanza | Dr. Phil	0.9384
	0.05	Modern Family | Shark Tank	Reba | The Young and the Restless	0.9299
	0.1	60 Minutes | Night Football	College Football | The Young and the Restless	0.9225
	0.2	NBA | Night Football	College Football | The Price Is Right	0.9244
	0.3	NBA | Night Football	College Football | CSI	0.9095
3	0.01	Modern Family | Bar Rescue | What Would You Do?	Duck Dynasty | Bonanza | Dr. Phil	0.9446
	0.05	CNN Anderson Cooper | Jimmy Kimmel | 60 Minutes	FNC Fox & Friends | The Young and the Restless | The First 48	0.9299
	0.1			0.9424

Table A2 (*continued*)

Max Size	Cut-off	Best chromosome Pro-Issue genes	Best chromosome Anti-Issue genes	Fitness
		Saturday Night Live | Kardashians | 60 Minutes	FNC Fox & Friends | College Football | The Young and the Restless	
	0.2	Shark Tank | Night Football | NBA	Hannity | College Football | The Price Is Right	0.9326
	0.3	CNN Anderson Cooper | Law & Order | Night Football	College Football | NBC Nightly News | CSI	0.9177
4	0.01	The Daily Show | Modern Family | Celebrity Family Feud | NBA	College Football | Duck Dynasty | Reba | 60 Days	0.9443
	0.05	CNN Anderson Cooper | 60 Minutes | Chopped | Shark Tank	Outnumbered | The Young and the Restless | The First 48 | Days of Our Lives	0.9411
	0.1	CNN Anderson Cooper | The Big Bang Theory | 60 Minutes | Shark Tank	NASCAR | Let's Make a Deal | Hawaii Five-0 | The Young and the Restless	0.9342
	0.2	NBA | Shark Tank | Blue Bloods | Night Football	FNC The Kelly File | College Football | The Price Is Right | The First 48	0.9380
	0.3	Saturday Night Live | NBA | 60 Minutes | Night Football	The O'Reilly Factor | College Football | 48 Hours | College Football	0.9188
5	0.01	The Big Bang Theory | Watch What Happens | Dancing with the Stars | This Is Us | Empire	Duck Dynasty | Bonanza | The Bold and The Beautiful | Scandal | Rosewood	0.9486
	0.05	CNN Anderson Cooper | Blue Bloods | Mom | Alaskan Bush People | Night Football	College Football | Reba | FNC The Five | Undercover Boss | Reba	0.9268
	0.1	CNN Anderson Cooper | The Big Bang Theory | Chopped | 60 Minutes | Naked and Afraid	FNC Fox & Friends | College Football | FNC The Five | The Young and the Restless | American Ninja Warrior	0.9420
	0.2	Saturday Night Live | Law & Order | Blue Bloods | Kardashians | 60 Minutes	College Football | The First 48 | CSI | Blue Bloods | 48 Hours	0.9312
	0.3	CNN Anderson Cooper | Saturday Night Live | NBA | The Big Bang Theory | Law & Order	College Football | American Pickers | CSI | NBC Nightly News | College Football	0.9219
6	0.01	The Good Place | Diners, Drive-Ins and Dives | NBA | Blacklist | What Would You Do? | Property Brothers	Gunsmoke | Rifleman | Charmed | College GameDay | The First 48 | UFC Fight	0.9478
	0.05	CNN Anderson Cooper | The Big Bang Theory | How I Met Your Mother | Shark Tank | 60 Minutes | Deadliest Catch	FNC Fox & Friends | The Price Is Right | American Idol | Reba | Days of Our Lives | American Pickers	0.9490
	0.1	CNN Anderson Cooper | Saturday Night Live | 60 Minutes | Super Bowl | Shark Tank | Diners, Drive-Ins and Dives	FNC Fox & Friends | College Football | American Dad! | American Ninja Warrior | The Young and the Restless | 48 Hours	0.9434
	0.2	CNN Anderson Cooper | Modern Family | NBA | Criminal Minds | Shark Tank | Night Football	College Football | The Price Is Right | The Walking Dead | Cops | The First 48 | CSI	0.9395
	0.3	CNN Anderson Cooper | Today Show | Night Football | NBA | Night Football | Property Brothers	College Football | NBC Nightly News | Forensic Files | 48 Hours | The Voice | College Football	0.92529

Table A3
Pro and Anti baskets for abortion rights.

Max Size	Cut-off	Best chromosome Pro-Issue genes	Best chromosome Anti-Issue genes	Fitness
2	0.01	Diners, Drive-Ins and Dives | The Good Place	Duck Dynasty | Bonanza	0.9246
	0.05	CNN Anderson Cooper | Shark Tank	Reba | The First 48	0.9146
	0.1	Saturday Night Live | Chopped	College Football | 48 Hours	0.9042
	0.2	CNN Anderson Cooper | Today Show	College Football | The Price Is Right	0.9043
	0.3	Today Show | NBA	College Football | NCIS	0.8938
3	0.01	CNN Anderson Cooper | Days of Our Lives | Million Dollar Listing	Bonanza | Reba | College GameDay	0.9291
	0.05	CNN Anderson Cooper | Madam Secretary | Diners, Drive-Ins and Dives	Reba | The First 48 | MasterChef	0.9097
	0.1	Madam Secretary | Today Show | Chopped	College Football | The Price Is Right | Let's Make a Deal	0.9157
	0.2	CNN Anderson Cooper | Saturday Night Live | Shark Tank	College Football | The Price Is Right | The Price Is Right	0.9081
	0.3	Today Show | 60 Minutes | NBA	College Football | CSI | NCIS	0.9002
4	0.01	MSNBC Rachel Maddow | Modern Family | Major Crimes | 60 Minutes	Reba | Rifleman | 48 Hours | The Bold and The Beautiful	0.9294
	0.05	Erin Burnett OutFront | The View | NBC Nightly News | Diners, Drive-Ins and Dives	College Football | NASCAR | Pawn Stars | Supernatural	0.9170
	0.1	CNN Anderson Cooper | 60 Minutes | CSI | Shark Tank	College Football | Let's Make a Deal | The First 48 | Naked and Afraid	0.9157
	0.2	SportsCenter | Shark Tank | America's Got Talent | 60 Minutes	Hannity | College Football | The Price Is Right | The First 48	0.9108
	0.3	60 Minutes | Good Morning America | NBA | Chopped	College Football | CSI | The Voice | 48 Hours	0.8970
5	0.01	The Good Place | Mike & Molly | Shades of Blue | Food Paradise | Diners, Drive-Ins and Dives	Bonanza | Rifleman | Chrisley Knows Best | Code Black | College GameDay	0.9262
	0.05	CNN Anderson Cooper | Cops | Jimmy Kimmel | Chopped | Property Brothers	College Football | Last Man Standing | CSI | Survivor | Person of Interest	0.9104
	0.1	CNN Anderson Cooper | Today Show | Criminal Minds | Madam Secretary | Kardashians	College Football | The Young and the Restless | The First 48 | CSI | 48 Hours	0.9185
	0.2	CNN Anderson Cooper | Today Show | Dancing with the Stars | Criminal Minds | Diners, Drive-Ins and Dives	College Football | The Price Is Right | Criminal Minds | Cops | The First 48	0.9124
	0.3	CNN Anderson Cooper | Law & Order | 60 Minutes | NFL Sunday | NBA	College Football | NCIS | Criminal Minds | 48 Hours | Good Morning America	0.9013
6	0.01	The Daily Show | Pioneer Woman | FNC For the Record w/ Greta | Leverage | Blindspot | Madam Secretary	Bonanza | Pit Bulls & Parolees | Lucifer | Rizzoli & Isles | Reba | The Bold and The Beautiful	0.92108
	0.05	CNN Anderson Cooper | Dancing with the Stars | Diners, Drive-Ins and Dives | Chopped | Shark Tank | Grey's Anatomy	FNC The Kelly File | Alaskan Bush People | 48 Hours | The Real Housewives | Reba | AFV	0.9190
	0.1	CNN Anderson Cooper | Modern Family | Let's Make a Deal | Blue Bloods | Diners, Drive-Ins and Dives | NBC Nightly News	College Football | NASCAR | The Young and the Restless | The First 48 | 20/20 | The Talk	0.9136
	0.2	CNN Anderson Cooper | Shark Tank | Today Show | Criminal Minds | Kardashians | 60 Minutes	College Football | The Price Is Right | Blue Bloods | 48 Hours | The First 48 | Diners, Drive-Ins and Dives	0.9154
	0.3	CNN Anderson Cooper | Chopped | 20/20 | Today Show | NBA | 60 Minutes	College Football | FNC The Kelly File | Fixer Upper | NBC Nightly News | 48 Hours | 20/20	0.9062

References

Blakley, J., Watson-Currie, E., Shin, H. S., Valenti, L. T., Saucier, C., & Boisvert, H. (2019). Are you what you watch? Tracking the political divide through TV preferences.

Bruell, A. (2019, June 04). Political Ad Spending Will Approach $10 Billion in 2020, New Forecast Predicts. Retrieved August 30, 2020, from https://www.wsj.com/articles/political-ad-spending-will-approach-10-billion-in-2020-new-forecast-predicts-11559642400.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Model. User-Adap. Inter., 12*(4), 331–370.

Chen, Y., & Yang, S. (2007). Estimating disaggregate models using aggregate data through augmentation of individual choice. *J. Mark. Res., 44*(4), 613–621.

Conover, M. D., Gonçalves, B., Ratkiewicz, J., Flammini, A., & Menczer, F. (2011). In *October). Predicting the political alignment of twitter users* (pp. 192–199). IEEE.

Cox, D., Lienesch, R., & Jones, R. P. (2017). Who sees discrimination? Attitudes on sexual orientation, gender identity, race, and immigration status—Findings from PRRI's American values atlas. PRRI. Available at www. prri. org/research/americans-views-discriminationimmigrants-blacks-lgbt-sex-marriage-immigration-reform.

Deng, Y., & Mela, C. F. (2018). TV viewing and advertising targeting. *J. Mark. Res., 55*(1), 99–118.

Emmerich, M., & Deutz, A. H. (2018). A tutorial on multiobjective optimization: Fundamentals and evolutionary methods. *Nat. Comput., 17*(3), 585–609.

Experian,. (2010). What Your TV Preferences Say About Your Politics. Retrieved September 04, 2020, from http://www.experian.com/blogs/marketing-forward/2010/11/15/what-your-tv-preferences-say-about-your-politics/.

Experian,. (2012). Top TV shows for reaching key voters. Retrieved January 9, 2020, from http://www.experian.com/blogs/marketing-forward/2012/08/28/top-tv-shows-for-reaching-key-voters/.

Goldberg, D. E., & Holland, J. H. (1988). Genetic algorithms and machine learning. *Machine learning, 3*(2), 95–99.

Hawkins, M. (2018). Eight TV Shows Conservatives Will Love. Retrieved January 9, 2020, from https://www.liveabout.com/conservative-tv-shows-4154645.

Hibberd, J. (2011, December 6). Republican vs. Democrat survey: Who watches the best TV shows? Retrieved January 9, 2020, from https://ew.com/article/2011/12/06/republican-vs-democrat-tv/.

Howe, P. D., Mildenberger, M., Marlon, J. R., & Leiserowitz, A. (2015). Geographic variation in opinions on climate change at state and local scales in the USA. *Nat. Clim. Change, 5*(6), 596.

Inan, E., Tekbacak, F., & Ozturk, C. (2018). Moreopt: A goal programming based movie recommender system. *Journal of computational science, 28*, 43–50.

Konak, A., Coit, D., & Smith, A. (2006). Multi-objective Optimization using Genetic Algorithms: A Tutorial. *Reliab. Eng. Syst. Saf., 91*, 992–1007.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl. Acad. Sci., 110*(15), 5802–5805.

Leonhardt, D. (2014, November 04). Football, the Newest Partisan Divide. Retrieved January 9, 2020, from https://www.nytimes.com/2014/11/04/upshot/football-the-newest-partisan-divide.html.

Liu, H. H., & Ong, C. S. (2008). Variable selection in clustering for marketing segmentation using genetic algorithms. *Expert Syst. Appl., 34*(1), 502–510.

Lovett, M. J. (2019). Empirical research on political marketing: A selected review. *Customer Needs and Solutions, 6*(3–4), 49–56.

Malthouse, E. C., Maslowska, E., & Franks, J. U. (2018). Understanding programmatic TV advertising. *International Journal of Advertising*, 1–16.

Meier, A. (Ed.). (2012). Fuzzy Methods for Customer Relationship Management and Marketing: Applications and Classifications: Applications and Classifications. IGI Global.

Musalem, A., Bradlow, E. T., & Raju, J. S. (2008). Who's got the coupon? Estimating consumer preferences and coupon usage from aggregate information. *J. Mark. Res., 45*(6), 715–730.

Musalem, A., Bradlow, E. T., & Raju, J. S. (2009). Bayesian estimation of random-coefficients choice models using aggregate data. *Journal of Applied Econometrics, 24*(3), 490–516.

NASA. (2019). Climate Change Evidence: How Do We Know?. Retrieved November 15, 2019, from https://climate.nasa.gov/evidence/.

Paine, N., Enten, H., & Jones-Rooy, A. (2017, September 29). How Every NFL Team's Fans Lean Politically. Retrieved January 9, 2020, from https://fivethirtyeight.com/features/how-every-nfl-teams-fans-lean-politically/.

Passwaiter, S., & Meininger, M. (2018). $5.25 Billion Spent During the Biggest Midterm Ad Blitz Ever. Retrieved September 9, 2020, from https://www.kantarmedia.com/us/thinking-and-resources/blog/5-25-billion-spent-during-the-biggest-midterm-ad-blitz-ever.

Public Religion Research Institute (n.d.). Retrieved November 13, 2019 from http://ava.prri.org/#lgbt/2016/States/lgbt_ssm/2,3.

Pew Research Center (n.d.). Views about abortion by state. Retrieved November 13, 2019 from https://www.pewforum.org/religious-landscape-study/compare/views-about-abortion/by/state/.

Rocha, M., & Neves, J. (1999). In *May). Preventing premature convergence to local optima in genetic algorithms via random offspring generation* (pp. 127–136). Berlin, Heidelberg: Springer.

Scott Diamond, J. (2014, October 6). TV's Most Republican and Democratic Shows. Retrieved January 9, 2020, from https://www.bloomberg.com/politics/graphics/2014-most-tv-campaign-ads/.

Sherkat, D. E., Powell-Williams, M., Maddox, G., & De Vries, K. M. (2011). Religion, politics, and support for same-sex marriage in the United States, 1988–2008. *Soc. Sci. Res., 40*(1), 167–180.

Tobias, M. (2018, March 22). Comparing Facebook data use by Obama, Cambridge Analytica. Retrieved January 9, 2020, from https://www.politifact.com/truth-o-meter/statements/2018/mar/22/meghan-mccain/comparing-facebook-data-use-obama-cambridge-analyt/.

Tuchman, A. E., Nair, H. S., & Gardete, P. M. (2018). Television ad-skipping, consumption complementarities and the consumer demand for advertising. *Quantitative Marketing and Economics, 16*(2), 111–174.

Vozalis, E., & Margaritis, K. G. (2003). September). Analysis of recommender systems algorithms. In *In The 6th Hellenic European Conference on Computer Mathematics & its Applications* (pp. 732–745).

Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proc. Natl. Acad. Sci., 112*(4), 1036–1040.