Traveling Canvassers

A. Matthew Eichinger

1 Introduction

Every election season more than a thousand politicians go door to door. They speak with constituents, ask about their troubles, and explain their policies. When they finish with one constituent, they move to the next. They have armies of activists doing the same. Eventually, they tire out and go home, exhaust the energy of their workers, or run out of time.

Canvassing is a staple of American elections. However, political scientists rarely study it. Why is unclear. Canvassing is the only action where politicians control almost all the terms of engagement. They choose to visit a specific constituent, pick a prepared speech from a set of speeches given what they know about the constituent, and predict follow up questions. Other campaign events are uncontrollable and impersonal. Town halls, stump speeches, debates, and television ads have uncertainty. You never know who will ask a question or what it will be; your clever debate opponent pokes holes in your logic that you could not foresee; your television audience varies. You can strategically account for these factors, but there may not be a good or practical way of doing so. With canvassing, you control the terms.

In this research note, I fill the theoretical gap. I first propose a theory for why I think canvassing happens. It revolves around "humanization and"quality information" - unique properties of face-to-face conversation. It also revolves around "base pumping" - visiting loyal patrons in electoral off-seasons to inspire future participation in on-seasons. I also propose an algorithm inspired by my theory for canvassing an arbitrary geographic space of heterogeneous constituents with an army of heterogeneous canvassers. The problem structure is similar to the traveling salesman but with refinements. You have multiple canvassers to assign; each has unique political, ethnic, and social skills; you know a good deal about your constituents; and you have a vote prediction machine - expert system, rule-based, parametric - you can query for the probability that a constituent votes for you before and after being canvassed.

A few insights are clear. First, politicians don't canvass if constituents have sticky imprints of their personalities. In this scenario, canvassing is a waste of time and money. Second, newer politicians have a bigger incentive to canvass. Information about them is scarce, so constituents have reason to assume the worst. Canvassing triangulates the information and makes it reliable. Third, never canvass extremists on the other side. No amount of engagement will change them. Extremists on your side will already vote for you, but a visit might get them to volunteer time for campaign events later in the season.

My research is useful - but is it "good?" Scholars might be uncomfortable with an algorithm that (1) creates an incentive to harvest voter data and (2) turn democratic elections into an optimization game. I get it, but I also have counters. (1) has been an incentive forever. My algorithm doesn't change that. (2) has always been the goal, but the best "game players" are rich and established politicians. My algorithm makes good performances more accessible to new and poor ones - it democratizes canvassing, so to speak. Moreover, canvassing is not the secret sauce to winning elections. It helps at the margins.

My research is also timely in that appears a little more than a year before the first US Presidential election in which foundational generative models will be present. Pundits are split on whether generative models will hurt, help, or do nothing to the integrity of election campaigns. The debate is about whether large language models, image generators, speech generators, and multimodal generators will be used nefariously to write expert propaganda and fool people into believing that candidates did something they did not do or said something they did not say. My theory of canvassing, though not directly addressing generative models, indirectly suggests we should not worry very much. Constituents already discount what they hear from politicians and are more inclined to listen to a face-to-face conversation. If constituents cannot trust the images and audio they see online, then they should discount it. This has the effect of raising the relative credibility of information they hear face-to-face, which they know has not been artificially generated.

My letter proceeds as follows. The first section discusses related literature. The second section introduces my theory. The third section introduces my algorithm, and my fourth and fifth sections discuss diagnostics and conclude.

2 Related Literature

Research on canvassing is divided into camps. Camp A studies electoral mobilization, or getting people to vote. Canvassing is treated as a tactic within electoral mobilization and gets minimal attention. Camp B studies canvassing itself. It mostly takes an historical and descriptive approach. For example, WIELHOUWER (1999) asked why and how American political parties in the South target voters, while Wielhouwer (2003), PANAGOPOULOS and WIELHOUWER (2008), and Beck and Heidemann (2014) discuss how grassroots campaign strategies are structured and change over time. They conclude that canvassing is highly strategic and serves multiple purposes, like collecting donations, creating volunteer lists, and establishing contact lists. Likewise, Chattharakul (2010) described canvassing in Thai elections in great detail. He narrowed in on meta tactics within the broader strategy of canvassing. For example, he found that Thai parties take pains to classify constituents people into homes or residencies, and to subsequently classify homes and residences into places that should or should not be contacted based on the expected political effect of a canvassing visit. His analytic framework was about deriving practical canvassing questions a political campaign needs to answer.

From this literature emerged three principles. First, visiting constituents serves multiple goals. It can raise the chance of voting, the chance of voting for you, the chance of donation, and the chance of participating in political events and organizations like rallies, public outreach, and non-profit work. Second, timing matters. Visiting constituents near election day mobilizes voters and persuades them to vote for you. Their minds are on elections because the information environment is dominated by political news and policy discussion. The consequence is that the marginal electoral utility of visiting constituents is (relatively) high just before an election and (relatively) low otherwise. This suggests that canvassing for non-electoral purposes should take place outside the time just before elections while canvassing for electoral purpose should take place just before elections. Third, having models of behavior, whether theoretical or empirical, is important. In a political campaign there should be at least three: (1) political participation as the outcome; (2) vote as the outcome; (3) vote for you as the outcome. Successfully canvassing requires (2) and (3) to be rigorous. For instance, some people don't want to be visited; visiting this type of person could persuade them to vote against you. Similarly, some people are racially prejudiced and respond poorly to conversations with members of a racial outgroup. To maximize the chance of winning an election, these factors must be accounted for.

A comprehensive theory of canvassing should incorporate all three principles. I do this in complicated way because I don't see an alternative. One model is for (non-electoral) political participation. Canvassing is an endogenous choice that raises the chance of participation. Resources are scarce, and depending on the quality of alternative tactics, canvassing will get more or less attention. The other model is for elections. In election season, you keep the canvassers from participation and you hire more if necessary. You choose whether to canvass for participation or voting based on their marginal utilities with respect to time. Since the marginal utility of electoral canvassing increases near elections - think of the marginal utility as being zero outside election seasons because voters forget - everyone gets assigned to canvassing duty.

I electoral theory of canvassing just needs to explain how canvassing should occur given that a politician has attracted some number of canvassers. Some could be hired, some could be operatives from other campaigns, and some could be people who were persuaded to participate from canvassing during the electoral off-season.

3 Theory

My premise is that conversing with a politician supplies a quality of information not obtainable elsewhere. Conversation reveals what the politician thinks, what the resident wants, and whether the politician understands the resident. A counterargument could be that since the resident knows the politician wants to fool them, they should discount what they hear in the conversation. This falls into the "cheap talk" category (Crawford and Sobel (1982)). There is some truth to this, but it is not determinative. Strategic residents, like partisan opponents, see elections as zero-sum and treat conversations as cheap talk. But non-strategic residents, like moderates and center-leaners, might see the conversation as a sincere attempt to hear what constituents think, compete for a job, and improve public service. There is no "correct" answer here - we are describing, rather than proscribing, behavior.

A Bayesian framework is useful. Suppose a resident has a prior belief about how much a politician would compromise in office. The variable is x and the prior belief is $x \sim N(\mu_0, \sigma_0^2)$. It makes sense to assume that prior beliefs come in three variants: (1) completely non-informative (politically uninformed); (2) tied to a common signal (general readers); and (3) tied to a common signal with bias (partisans). I model this by assuming that μ_x is an outcome variable. In the non-informative case, μ_0 is $\mu_0 = 0$. In the common signal case, $\mu_0 = a$, where $a \in R$ is how the media portrays the politician. In the common signal with bias case, $\mu_0 = a + b_g$, where $b_g \in R$ is the bias for partisan group g. The same forensic analysis of the variance term is worthwhile. I make it simple by noting that partisans are more rigid than general readers, and general readers are more rigid than non-informed people. In terms of the model, this means that partisans have the smallest spread around their belief, general readers have the middle spread, and the non-informed have the largest spread.

Suppose a resident gets visited by a politician. He starts with a prior belief. As the conversation goes on, he learns more about the politician. They get a better signal of what the politician believes, how much she knows about policy and public needs, and how much she would compromise her ideology to get things done. As the conversation finishes, the resident updates his belief. Using the equation for Bayesian updates, the new mean belief is a weighted average of information from the conversation and information from the prior:

$$\mu_1 = \mu_0 + (x - \mu_0)\frac{\sigma_0^2}{\sigma_0^2 + \sigma_x^2}, \tag{1}$$

where σ_x^2 is the variance of the new information. The variance of μ_1 is then

$$\sigma_1^2 = \frac{\sigma_0^2 \sigma_x^2}{\sigma_0^2 + \sigma_x^2}$$

Easy principles emerge from this setup. The first is evasion of partisan opponents. Partisan opponents have precise and rigid prior beliefs. This implies σ_0^2 is small. Per equation (1), the consequence is that partisan residents rely on their prior σ_0 more than what they hear in the conversation x. The second principle is that non-informed people are the easiest targets. Their prior beliefs are spread wide. This implies σ_0^2 is large. A conversation to learn x would get weighed highly. The third principle is that informed but non-partisan people are good but not great targets. This has to do with $\mu_a = 0$, and the fact that a could be hard to move.

4 Algorithm

The theory section says the electoral value of canvassing is disseminating information to people who would count it as credible. You stay away from extreme partisan opponents and extreme partisan loyalists. Opponents discount what you say; loyalists already agree. You want to target people who haven't made up their mind and people who need a little nudge to vote.

We can use this idea to build our computational model of canvassing. The model has four components: (1) geographic representation of residents, (2) generating descriptive characteristics for canvassers, residents, and politicians, as well as specifying predictive models for how people vote, (3) making initial assignments of canvassers to residents, and (4) making assignments after the initial assignment. The algorithmic portion of the steps is (4); the other steps

4.1 Geography

Four elements comprise the algorithm. The first is a geographical element. I assume constituents are located on a two-dimensional xy plane. The boundaries of the plane are $(x_{\min}, x_{\max})$ and $(y_{\min}, y_{\max})$. A resident $r \in 1, 2, ..., R$ lives at $\boldsymbol{z}_r = [x_r, y_r]'$, which corresponds to Cartesian coordinates. The absolute distance between a pair of arbitrary residents (i, j) is therefore

$$\text{travel time}_{ij} = d(\boldsymbol{z}_i, \boldsymbol{z}_j) = |\boldsymbol{z}_i - \boldsymbol{z}_j| = |x_i - x_j| + |y_i - y_j|. \tag{2}$$

Travel distances can be stored in an $R \times R$ adjacency matrix A, where element $a_{ij} = d(\boldsymbol{z}_i, \boldsymbol{z}_j)$. I use this shorthand notation a_{ij} for travel distances for the remaining discussion.

Geographic distributions of people in the real world are clustered. Residents live close to each other in discrete blocks of (seemingly) partitioned spaces. To represent this, I generate B neighborhoods $b \in 1, 2, ..., B$. Each neighborhood b has a central anchor at position (x_b, y_b). I draw anchors from a random uniform distribution with components $x_b \sim U(x_{\min}, x_{\max})$. Then, for each neighborhood $b \in B$, I randomly draw x and y offsets for K residents $k \in 1, 2, ..., K$ from a random normal distribution with mean $\mu_x = \mu_y$ and standard deviation $\sigma_x = \sigma_y$.

Figure 1 shows how this looks with two examples. For this particular case, I set $B = 50$ and $K = 400$, which makes $R = 20,000$ residents. The red dots indicate neighborhood anchors and the black dots indicate resident positions.

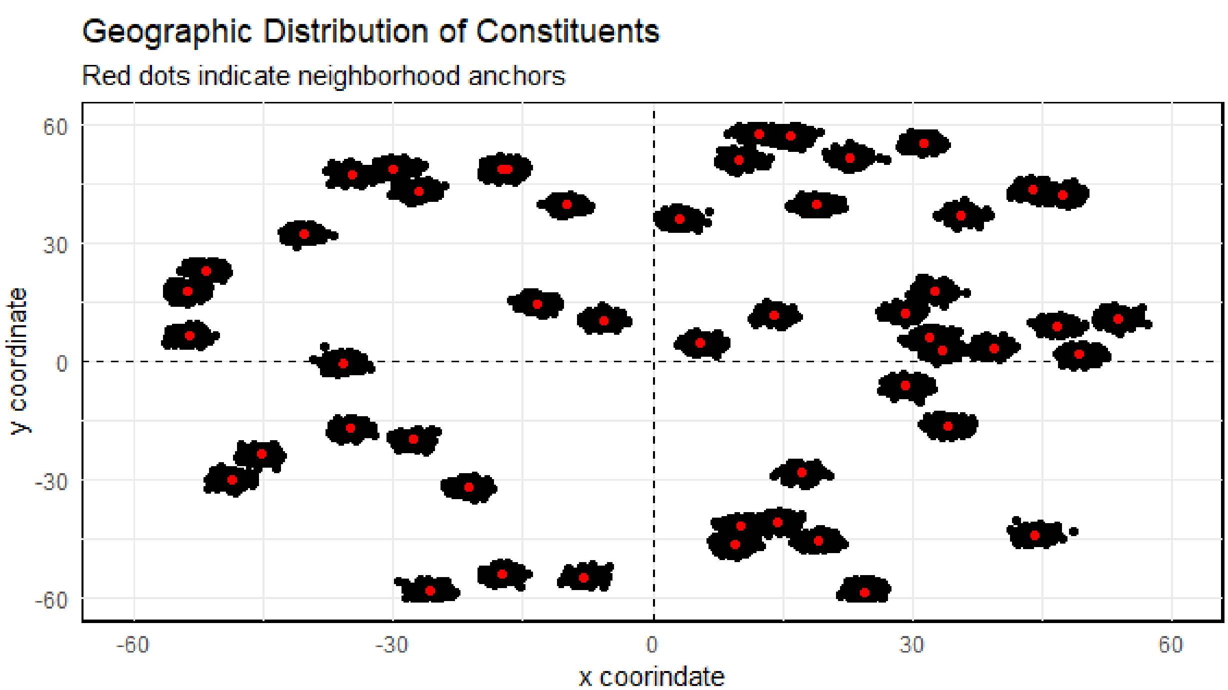

Figure 1: Insert caption.

4.2 Characteristics and Predictive Models

The second elements are the predictive models. There is a vote model to predict voting $v_r \in \{0, 1\}$ and a choice model to predict candidate choice $Y_r|V_r \in \{0, 1\}$. I denote $V_r = 1$ as voting and $Y_r = 1$ as voting for the candidate of interest.

The predictive models rely on canvasser and resident data. You have C canvassers $c \in 1, 2, ..., C$ Each has physical and sociodemographic characteristics like race, ethnicity, height, weight, language, age, income, and education. Each also has conversation skills, political expertise, and ideological versatility, which makes him or her better able to speak to partisans. Without loss of generality, there are P features that describe your canvassers $\boldsymbol{x}_c = \{x_{c1}, x_{c2}, ..., x_{cP}\}$. You collect the information yourself.

The third element is residents. There are $R = B \times K$ residents, who are indexed by $r \in 1, 2, ..., R$. Each resident has characteristics like the canvassers but most may be masked due to data limitations. You only know about residents you have data on. For example, you might know party affiliation, vote history, income bracket, race, and nationality. There are Q features that describe the residents $\boldsymbol{x}_r = \{x_{r1}, x_{r2}, ..., x_{rQ}\}$. You need to purchase or scrape the information.

How you leverage data in predictive models is arbitrary. The models can be expert based or quantitative; parametric or non-parametric; regularized or non-regularized; a single model or an ensemble. The only rule is that they be able to query or predict the change in vote chance and vote choice after being visited by canvasser c with characteristics $\boldsymbol{x}_c$. Let the visit indicator be $g \in \{0, 1\}$. You need to be able to compute the change in vote chance

$$\Delta V(\boldsymbol{x}_r, \boldsymbol{x}_c, g) = \Pr(V_r = 1|\boldsymbol{x}_r, \boldsymbol{x}_c, g = 1) - \Pr(V_r = 1|\boldsymbol{x}_r, \boldsymbol{x}_c, g = 0) \tag{3}$$

and the change in vote choice

$$\Delta Y(\boldsymbol{x}_r, \boldsymbol{x}_c, g) = \Pr(Y_r = 1|\boldsymbol{x}_r, \boldsymbol{x}_c, g = 1) - \Pr(Y_r = 1|\boldsymbol{x}_r, \boldsymbol{x}_c, g = 0). \tag{4}$$

Due to resident tastes over race, ethnicity, nationality, social characteristics, and politics, some resident-canvasser visits could lead to worse election chances. For example, a highly conservative resident may only want to be visited by a canvasser who shares their race. Neglecting this factor but making the visit nonetheless could make a resident both likelier to vote and likelier to vote against you. In this case, ΔV would be greater than and ΔY would be less than zero.

4.3 Initial Assignments

Canvassers must start their journey from somewhere. But where? Is there an optimal start? The short answer is no. In computer science terms, the problem is NP-hard. That means there is an optimal solution but no analytic way (e.g., backwards induction) to get there. You have to try every combination of starting position and path trajectory, calculate expected vote gains, and pick the highest scoring combination. The core problem is that sending canvasser c to resident r removes r from the table and commits c to a geographic position. You never know if r would have been on a different canvasser's path later on, or if c should have waited to visit r when c was in a different position. Consequently, there is no obvious method of initial assignment.

The best course of action is to use heuristic starts. Three are intuitive: random assignment to residents, multi-person best assignment, and single-person best assignment. Random assignment samples a resident without replacement for each canvasser. Multi-person best assignment picks the best resident to visit for canvasser given their current location, checks for (and resolves) matching conflicts, and makes C assignments. Single-person best assignment just picks the single best resident-canvasser pair on the board and makes the assignment; it is a slower and more cautious version of multi-person assignment. In this framework, "best" assignments are determined using equations (3) and (4). These methods are good starting points but their performances must be numerically tested.

After a method of initial assignment is chosen, the locations of canvassers should be stored and updated. Call the storage object the "tracker" $\boldsymbol{T}$. The tracker is a vector $\boldsymbol{T} = \{t_1, t_2, ..., t_C\}$, where t_c denotes the location of canvasser c using resident locations $\boldsymbol{z}_r$. In other words, $t_c \in \boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_R$. When canvasser c gets assigned to resident d, the tracker gets updated such that $t_c = \boldsymbol{z}_d$.

4.4 Optimization

After initial assignments are made, how are canvassers allocated to residents? Recall that the objective is to use canvassing to maximize expected number of votes. Canvassers should be sent to residents where they'll have the highest vote effects. These are already stored in the $R \times C \times M$ lookup table L. We want to find resident-canvasser combinations with the highest vote effects and send our canvassers there.

The trick and constraint are traveling time. If there was no travel time, we could send canvassers from neighborhood to neighborhood, in and out of apartments, and into the paths of canvassers without concern to the residents they impact the most. However, canvassers have limited time. Their assignments have to be strategic. By sending a canvasser to residents they would have the highest effect on, they might end up crisscrossing the space and reaching only a handful of people. It might be better to stay in a neighborhood and visit a high number of residents they would have a moderate effect on. The question is how to overcome travel time.

A good way to account for travel time is a decision framework. At an arbitrary spot $\boldsymbol{z}_c = t_c$, canvasser c can choose to go anywhere. By moving to resident r, they gain would gain $\Delta V(\boldsymbol{x}_r, \boldsymbol{x}_c = t_c, g)$ and $\Delta Y(\boldsymbol{x}_r, \boldsymbol{x}_c = t_c, g)$ at travel cost a_{ij}. On the flip side, they would give $1 - \Delta V(\boldsymbol{x}_r, \boldsymbol{x}_c, g)$ and $1 - \Delta Y(\boldsymbol{x}_r, \boldsymbol{x}_c, g)$ to the other candidate. These terms form the "vote swing" expression

$$EU(g = 1) \geq EU(g = 0)$$
$$\text{what you get}(g = 1) - \text{what you give}(g = 1) \geq \text{what you get}(g = 0) - \text{what you give}(g = 0)$$
$$\Pr(V_r = 1|\boldsymbol{x}_r, \boldsymbol{x}_c, g = 1)W(\boldsymbol{x}_r, \boldsymbol{x}_c, g = 1) \geq \Pr(V_r = 1|\boldsymbol{x}_r, \boldsymbol{x}_c, g = 0)W(\boldsymbol{x}_r, \boldsymbol{x}_c, g = 0),$$

where $W(\boldsymbol{x}_r, \boldsymbol{x}_c, g = 1) = \Pr(Y_r = 1|\boldsymbol{x}_r, \boldsymbol{x}_c, g = 1) - \Pr(Y_r = 0|\boldsymbol{x}_r, \boldsymbol{x}_c, g = 1)$ is the difference in vote choice probabilities for candidates when visiting and $W(\boldsymbol{x}_r, \boldsymbol{x}_c, g = 0) = \Pr(Y_r = 1|\boldsymbol{x}_r, \boldsymbol{x}_c, g = 0) - \Pr(Y_r = 0|\boldsymbol{x}_r, \boldsymbol{x}_c, g = 0)$ is the difference in vote choice probabilities when not visiting. The left-hand side describes the total chance that resident r votes given their features $\boldsymbol{x}_r$, the features of the canvasser $\boldsymbol{x}_c$, and that they receive a visit $g = 1$; the right-hand side describes the total chance given they do not. By re-arranging this expression and accounting for travel time a_{ij}, we get the equation for the net vote swing:

$$\text{Net vote swing} = S(\boldsymbol{x}_r, \boldsymbol{x}_c, g) = \frac{1}{a_{ij}} \left(\frac{\Pr(V_r = 1 | \boldsymbol{x}_r, \boldsymbol{x}_c, g = 1)}{\Pr(V_r = 1 | \boldsymbol{x}_r, \boldsymbol{x}_c, g = 0)} \frac{W(\boldsymbol{x}_r, \boldsymbol{x}_c, g = 1)}{W(\boldsymbol{x}_r, \boldsymbol{x}_c, g = 0)} - 1 \right), \quad (5)$$

where a_{ij} is a penalizing factor from equation (2) accounts for time spent traveling from t_c to r. When equation (5) is above zero, it means the total effect of canvassing is in your favor (i.e., the "vote swing" is in your favor). The size of the effect denotes how much. When equation (5) is below zero, it means the total effect is in the favor of your opponent. In this case, you *do not* have an incentive to canvass. It would benefit your opponent more than you and would therefore hurt your election prospects.

Given locations $\boldsymbol{T}$, equation (5) can be applied to every resident-canvasser pair. It completely describes the rewards on the board, conditional on locations. When used by canvasser c on every active resident $R\prime = \{r : r \in R \wedge U_r = 0\}$, it produces a vector of expected rewards $\boldsymbol{e}_c = \{S(\boldsymbol{x}_1, \boldsymbol{x}_c, g), S(\boldsymbol{x}_2, \boldsymbol{x}_c, g), ..., S(\boldsymbol{x}_{R\prime}, \boldsymbol{x}_c, g)\}$. This vector can be rank-ordered into $\boldsymbol{e}_c\prime$, which can be queried for rank b like $\boldsymbol{e}_c\prime[b]$. For example, the first-ranked element would be $\boldsymbol{e}_c\prime[1] = \max(\boldsymbol{e}_c)$. By applying this method to every canvasser $c \in c$, we get the set of all possible visits $\boldsymbol{F} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, ..., \boldsymbol{e}_C\}$. Then, by rank-ordering F and removing elements for which equation (5) is below zero, we get the "feasible set" of visits

$$\text{Feasible set} = \boldsymbol{F}\prime = \{\boldsymbol{F} : S_1(\boldsymbol{x}_r, \boldsymbol{x}_c, g) > S_2(\boldsymbol{x}_r, \boldsymbol{x}_c, g) \ \wedge \ S(\boldsymbol{x}_r, \boldsymbol{x}_c, g) > 0\}.$$

where S_1 is the first-best visit, S_2 is the second-best visit, and so forth. The condition $S(\boldsymbol{x}_r, \boldsymbol{x}_c, g) > 0$ is the rule that elements must be greater than 0.

The final step is choosing an assignment framework. From the NP nature of the problem, we only have heuristics. The first framework is a greedy algorithm. Given the tracking vector $\boldsymbol{T}$, assign canvassers based on vote swing. There are two meta methods. Meta method (a) is conservative and assigns the single canvasser with the highest vote swing on the board to the corresponding resident r; it updates the tracking vector by removing r from the choice set, removing r from the lookup table L, and updating the feasible visits $\boldsymbol{F}\prime$ through equation (5) given c's new position at $\boldsymbol{z}_r$. Meta method (b) uses the same operation but assigns each canvasser their highest vote swing. If there are ties (i.e., two canvassers c and d have resident r as their highest vote swing), they are broken by picking which canvasser c or d gets the higher vote swing. The tracking vector, lookup table, and feasible visits are updated by removing the chosen residents and adjusting a_{ij} for new positions. This meta method is less conservative in that it favors speed over performance. Suppose the next two steps of meta method (a) would assign canvasser c to resident r and $r\prime$. These are the best moves on the board. What happens if meta method (b), in making C assignments at once, assigns a canvasser $\neg c$ to $r\prime$? We know this is non-optimal from looking at the first two assignments of meta method (a). Had meta method (b) been patient, it could have earned a higher reward by keeping $r\prime$ for c; but patience comes with compute time, and this could be prohibitive.

The steps of the algorithm are shown in box (). The steps are organized into premises and a loop. The premises describe what information we have beforehand and how the information is organized into data structures. For example, box() explains that we have fitted models for voting $\hat{V}(\boldsymbol{x}_r, \boldsymbol{x}_c, g)$ and vote choice $\hat{Y}(\boldsymbol{x}_r, \boldsymbol{x}_c, g)$, a tracking vector $\boldsymbol{T}$, an adjacency matrix $\boldsymbol{A}$, vote swing equation S, and a feasible assignments vector $\boldsymbol{F\prime}$. These objects contain every piece of information necessary for the algorithm to operate. In our case above, The loop section explains how this information is used to assign canvassers to residents for maximizing electoral performance. It goes into the single-person method, the multi-person method, and the random method. The methods are similar in how they update tracking information across iterations but different in their criteria for assignments.

5 Diagnostics

I ran each of the three methods ten times. For each run i, I recorded the number of expected votes gained as $G(m_i)$. I took the average of this performance measure for a method $\frac{1}{10}\sum_i^{10} G(m_i)$ to gauge how well it performed. Figure @(ref:perf) shows this using density plots. It ignores the "random" method because it performed so poorly. The "fast" method is indicated by a solid line while the "slow" method is indicated by a dashed line. The densities suggest the slow single-person method performs slightly better than the fast multi-person method. On average, the fast method gained 264.45 expected votes, where the slow method gained 274.86. This is a small difference in light of their speed difference. The fast method finished 10 runs in less than twenty seconds while the slow method took more than a thirty minutes.

How do the optimal paths look? Figures 3 and 4 plot path trajectories for single-person and random methods, respectively. The multi-person method looks tremendously similar to the single-person method, so I left it out. In the figures, the line colors indicate canvasers. As expected, trajectories in the random panel 4 are terrible. Canvassers move all over and waste time. By contrast, the slow method is sensible 3. Most canvassers stay in a neighborhood, while a small group moves across neighborhoods. This was predictable because there is a large incentive to make visits with small travel times.

The path trajectories reveal that a neighborhood heuristic is reasonable. In the slow method, canvassers mostly stick to neighborhoods. The majority of canvassers stay in the neighborhood they were first assigned. These are "fine" canvassers who are good but not great at convincing people to vote. A small minority of canvassers move from neighborhood to neighborhood. These are "great" canvassers who are the most persuasive conversationalists. This pattern leads to a couple heuristics:

- Assign canvassers to distinct neighborhoods;
- If possible, give your best canvassers multiple neighborhoods.

6 Conclusion

I introduced a basic theory of canvassing, along with a canvassing algorithm based on theoretical insights. The components of the algorithm should be familiar to political scientists

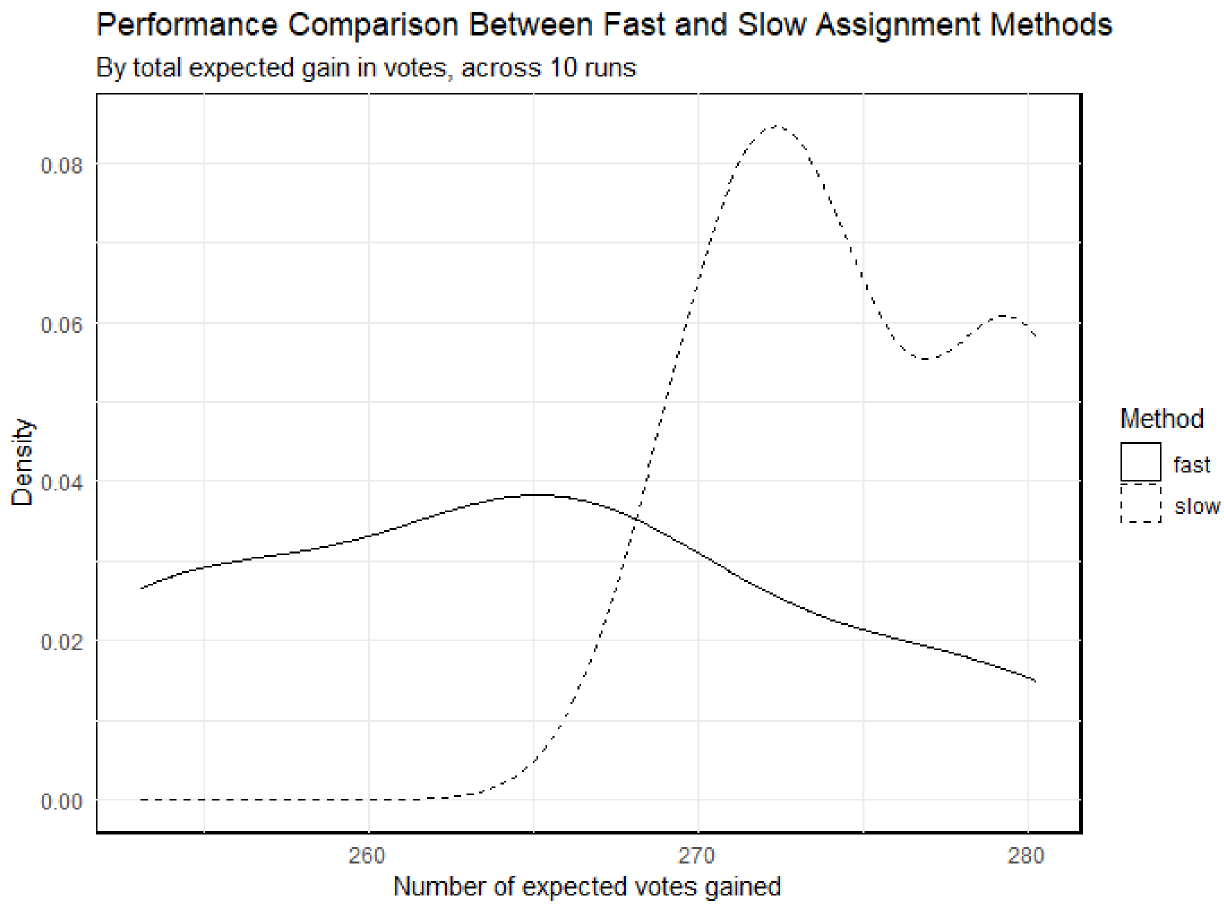

Figure 2: Blop.

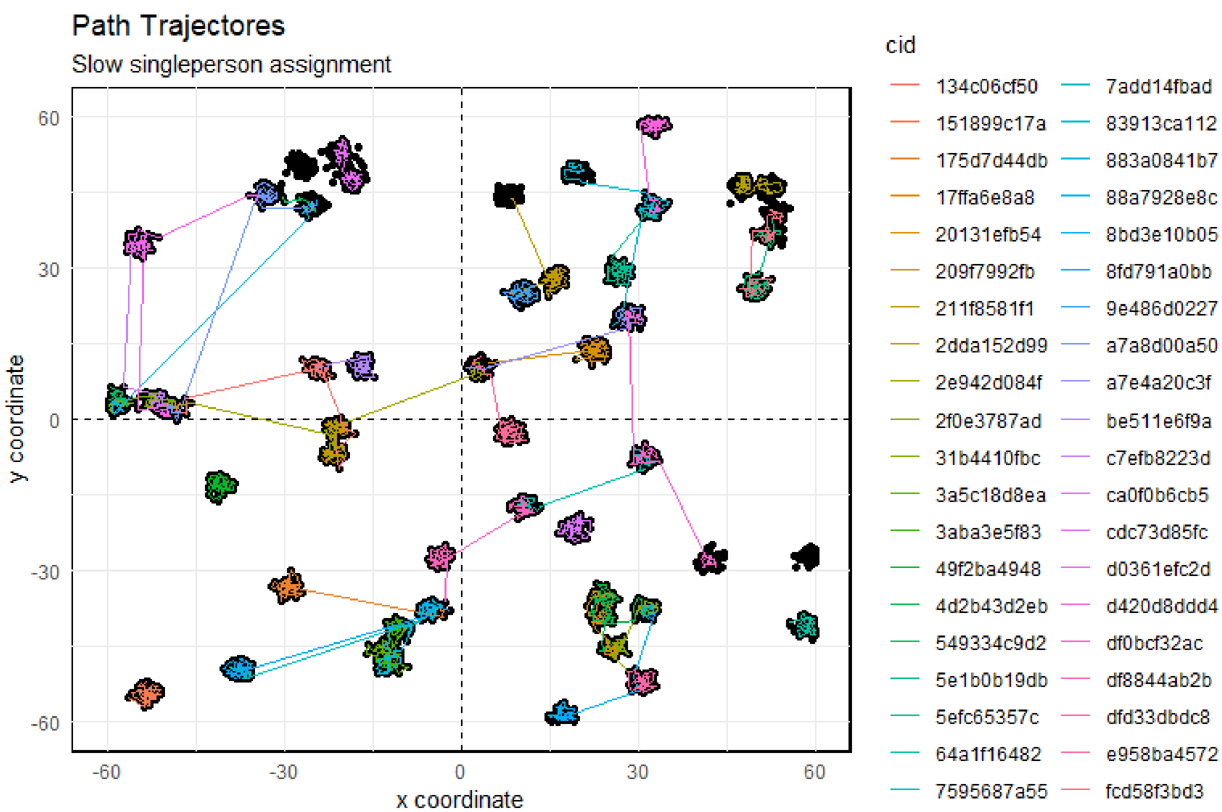

Figure 3: Blop.

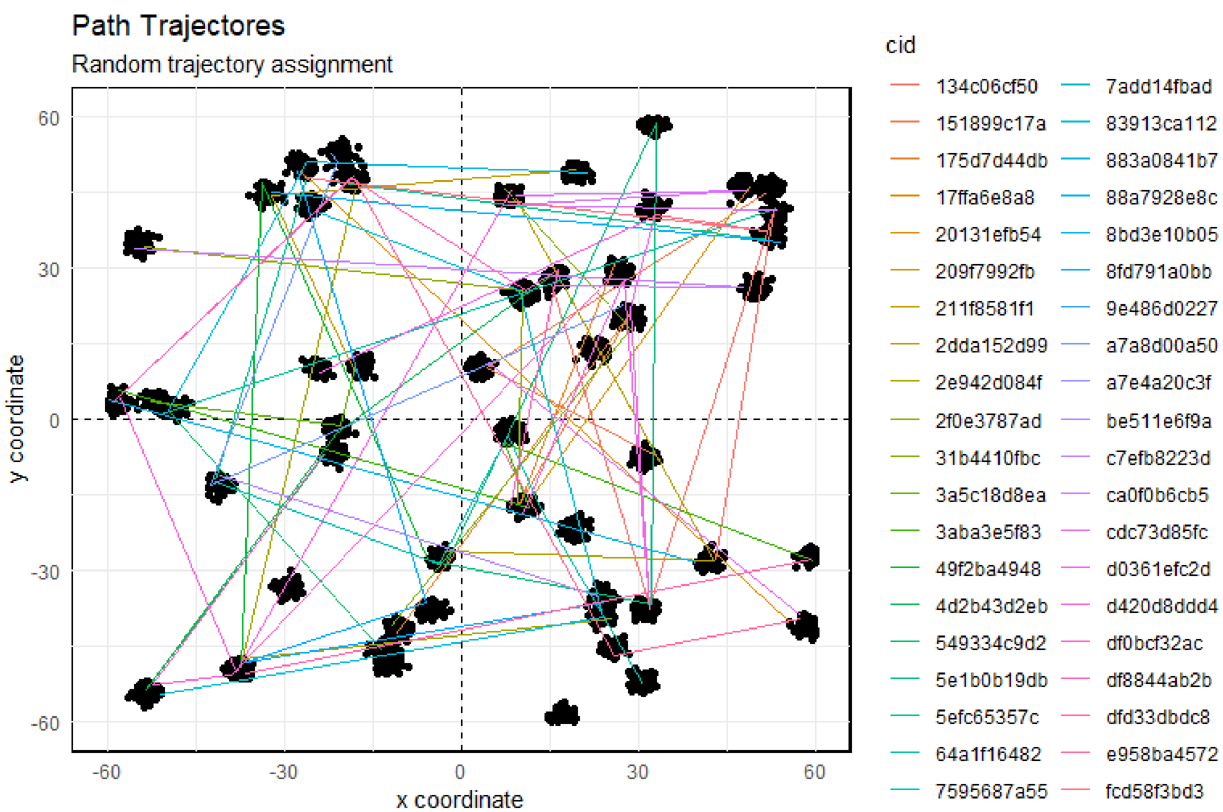

Figure 4: Blop.

because it relies heavily on political and electoral theory. They should also be familiar to computer scientists because they are similar to elements of the traveling salesman problem.

What could be improved? I have a couple things in mind. The first is to model social externalities. If you have a predictive model for "spreader residents," or people who tell their friends they received a visit, then you could include this in a model. In essence, a spreader raises the chance that people in their neighborhood vote for you by $0 \leq k \leq 1$ probability units. Most likely, $k \approx 0$, and the data needed to identify spreaders might be expensive or unobtainable.

The second improvement is to model apartment units. I chose not to model apartments because many apartments may not approve political solicitation. Adding apartments to the algorithm is "easy" because they are just dense geographic areas where travel time between residents is small. However, it is also "hard" because it adds a vertical dimension to the geography. I think the best way to handle apartments is to add small if-else rule: if a canvasser enters an apartment from the outside, add K minutes for dealing entrance hassles; if a canvasser leaves an apartment, add $K - L$ minutes for finding the exit; else, keep travel times. This condition should be imposed on the Q matrix when adjusting best visits for travel times based on canvasser positions.

The final extension is to model social relations. In theory, there is a giant $R \times R$ social relations matrix B describing how socially close residents are. Each element $b_{ij} \in \{0, 1\}$ describes how close i and j are. Taking from sociology and psychology, we know the choices of an individual i depend on their preferences and the choices of people around them. We also know people select into social groups where members are similar to themselves, a term known as homophily (a.k.a., assortative matching, assortative mating). The consequence is that if we know B, then we can make intentional visits to an individual i in the hopes of influencing their friends $J_i : b_{ij} > 0$. This would be implemented in the algorithm in two steps. The first would be to calculate another expected value from a visit: a social transmission effect

$$\begin{aligned}\text{Social transmission effect} = \sum_{j \in J_i} & Y(V_j = 1 | W_i = 1, b_{ij}) - Y(V_j = 1 | W_i = 0, b_{ij}) \cdot \\ & Y(U_j = 1 | V_j = 1, W_i = 1, b_{ij}) - Y(U_j | V_j = 1, W_i = 1, b_{ij})\end{aligned}$$

The nasty expression has two parts. The term inside the brackets is the change in the probability of voting for person j when their friend i gets visited. The difference is conditioned by b_{ij}, or i and j's closeness. When $b_{ij} = 0$, the term goes to zero and the social transmission effect from i to j is zero. The term getting multiplied by the bracketed term is the change in the probability that j votes for the politician when i gets visited.

Capturing social transmission effects is likely not worth the work and price of data. Since people select themselves into social relations, and are consequently influenced by their friends independent of a visit, the social transmission affect probably contributes nothing. Other variables in the predictive model, those in $\boldsymbol{x}_r$, do the work. The social transmission idea is interesting but not worth the time.

7 References

Beck, Paul A, and Erik D Heidemann. 2014. "Changing Strategies in Grassroots Canvassing." *Party Politics* 20 (March): 261–74. https://doi.org/10.1177/1354068813509518.

Chattharakul, Anyarat. 2010. "Thai Electoral Campaigning: Vote-Canvassing Networks and Hybrid Voting." *Journal of Current Southeast Asian Affairs* 29 (December): 67–95. https://doi.org/10.1177/186810341002900404.

Crawford, Vincent P., and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50 (November): 1431. https://doi.org/10.2307/1913390.

Huckfeldt, Robert, and John Sprague. 1992. "Political Parties and Electoral Mobilization: Political Structure, Social Structure, and the Party Canvass." *American Political Science Review* 86 (March): 70–86. https://doi.org/10.2307/1964016.

PANAGOPOULOS, COSTAS, and PETER W. WIELHOUWER. 2008. "Polls and Elections the Ground War 2000-2004: Strategic Targeting in Grassroots Campaigns." *Presidential Studies Quarterly* 38 (June): 347–62. https://doi.org/10.1111/j.1741-5705.2008.02645.x.

WIELHOUWER, PETER W. 1999. "The Mobilization of Campaign Activists by the Party Canvass." *American Politics Quarterly* 27 (April): 177–200. https://doi.org/10.1177/1532673X99027002002.

Wielhouwer, Peter W. 2003. "In Search of Lincoln's Perfect List." *American Politics Research* 31 (November): 632–69. https://doi.org/10.1177/1532673X03255171.