Université de Montréal

Fair Vaccination Strategies with Influence Maximization: A Case Study on COVID-19

par

Nicola Neophytou

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Informatique

November 8, 2023

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

Fair Vaccination Strategies with Influence Maximization: A Case Study on COVID-19

présenté par

Nicola Neophytou

a été évalué par un jury composé des personnes suivantes :

Pierre-Luc Bacon
(président-rapporteur)

Golnoosh Farnadi
(directeur de recherche)

Guillaume Rabusseau
(membre du jury)

Résumé

Pendant la pandémie de Covid-19, les minorités raciales et les groupes économiquement défavorisés ont connu des taux accrus d'infection, d'hospitalisation et de décès dans les zones urbaines. Cette disparité témoigne de l'oppression systématique à laquelle sont confrontées les minorités raciales et la classe ouvrière, qui s'étend évidemment aux services de santé. Les inégalités flagrantes en matière de santé étaient évidentes avant que les vaccins ne soient disponibles, nous ne pouvons donc pas simplement les attribuer à des attitudes culturelles d'hésitation à la vaccination. Dans ce travail, nous présentons des solutions pour optimiser la distribution équitable des vaccins pour différents groupes démographiques, afin de promouvoir un accès équitable aux vaccins lors du premier cycle d'attribution. Nous nous appuyons sur des travaux antérieurs pour construire des réseaux de mobilité de trois zones métropolitaines américaines en utilisant des données de visites réelles dans des lieux publics au cours des premières semaines de la pandémie. Nous proposons une nouvelle méthode utilisant la maximisation de l'influence pour détecter les quartiers les plus influents de la zone urbaine en termes d'efficacité dans la propagation de la maladie. Nous modélisons ensuite la propagation ultérieure de la maladie avec ces quartiers sélectionnés vaccinés. De plus, nous introduisons des considérations d'équité afin de mettre en œuvre un accès équitable aux vaccins pour les groupes raciaux et les groupes de revenus du réseau. Pour fusionner nos solutions avec les stratégies actuelles, nous combinons nos stratégies équitables avec une méthode de priorisation pour les groupes plus âgés du réseau.

Mots clés: Maximisation de l'influence, réseaux de mobilité, distribution de vaccins, équité démographique

Abstract

During the Covid-19 pandemic, racial minorities and economically-disadvantaged groups experienced heightened rates of infection, hospitalization and death in urban areas. This disparity speaks to the systematic oppression faced by racial minorities and the working classes, which evidently extends to healthcare provisions. The stark inequalities in health outcomes were clear before vaccines became available, so we cannot simply attribute this to cultural attitudes of vaccine hesitancy. In this work, we present solutions to optimize the fair distribution of vaccines for different demographic groups, in order to promote equitable vaccine access in the first round of allocation. We build on previous work to construct mobility networks of three US metropolitan areas using data of real visits to public places during the first weeks of the pandemic. We propose a novel method using influence maximization (IM) to detect the most influential neighborhoods in the urban area in terms of efficacy in spreading the disease. We then model the subsequent disease spread with these selected neighborhoods vaccinated. Additionally, we introduce fairness considerations, to implement equitable vaccine access for racial groups and income groups in the network. To merge our solutions with current strategies, we combine our fair strategies with a prioritization method for older-age groups in the network.

Keywords: Influence Maximization, mobility networks, vaccine distribution, demographic fairness

Contents

List of tables

List of figures

List of acronyms and abbreviations

AI	Artificial Intelligence
IM	Influence Maximization
CBG	Census Block Group - geographical divisions of US neighborhoods
POI	Points of Interest - public places such as gyms, restaurants etc.
MSA	Metropolitan Statistical Area - geographical regions in the US with high population density, containing several cities and major towns
SEIR	Susceptible, Exposed, Infected, Recovered - both a type of epidemiological model and reference to the possible infection stages
CELF	Cost-Effective Lazy Forward selection - an Influence Maximization technique [**50**]

Acknowledgements

I'd firstly like to thank the jury for dedicating their time to evaluate this thesis. I sincerely hope you enjoy it!

I dedicate this thesis to my loving parents and wonderful sister, for their unwavering support throughout this Masters, and in every step that came before (in particular for supporting my move across the Atlantic).

I owe significant thanks to Afaf Taïk and Professor Golnoosh Farnadi for their passion, consistency and extremely valuable guidance.

Very special thanks to Rebecca Salganik, Kiarash Mohammadi and Rohan Sukumaran for their solidarity, faith, encouragement, optimism and food. You have been more instrumental than you know.

Finally, an extremely important thank you to Thomas Holweg-Bruneau, who has inspired me everyday.

Chapter 1

Introduction

In the outbreak of the Covid-19 pandemic, economically-disadvantaged groups and marginalized racial groups experienced higher rates of infection, hospitalization and death [**82, 70, 75, 64, 40, 8, 36, 30**]. These effects are particularly striking in urban areas of Western countries; in Chicago, African-Americans make up 30% of the population, yet in 2020 they represented 50% of Covid-19 cases and 70% of Covid-19 deaths [**77**]. In Louisiana, they represented 32.2% of the population and 70.5% of Covid-19 deaths. In New York City, the urban area with the largest counts of cases and deaths in the country, African-Americans and Hispanics/Latinos accounted for 22% and 29% of the population, and represented 28% and 34% of deaths respectively [**2**]. The sobering statistics signal the extreme ways in which deprived communities in Western urban areas still benefit disproportionately less from health and social care resources compared to the economically-favoured in the same areas. This inequality in access to healthcare was also reflected when vaccines became available; vaccine uptake rates were lower amongst communities of color in the US and the UK [**9, 74**]. It has long been the case, as exhibited in other disease outbreaks before Covid-19, that minorities are failed in three major ways: disparities in exposure to the virus, disparities in susceptibility to contracting the virus, and disparities in treatment [**12**].

In the US, CDC data revealed that in April 2021, the gap in vaccination rates across all states between White and Black groups was a 14 percentage point difference, and between White and Hispanic groups was 13 percentage points [**69**]. Similarly, a study of individuals aged fifty and over in Wales, UK found that the likelihood of being vaccinated was for lower for males, non-whites, and people living in deprived areas and urban areas [**74**]. The largest inequality was observed between racial groups, with the gap between average rate of vaccination being 20.2 percentage points higher amongst those identifying as White than those identifying as Black. They also revealed that this gap widened over the first five

months of the vaccine roll-out, as other additional priority groups became eligible to receive vaccination.

We cannot simply attribute this discrepancy to the difference in vaccine hesitancy due to cultural attitudes (though this is often the focus of much of the literature [**23, 85**]); we must also consider additional barriers that economically-deprived communities must overcome in order to get vaccinated. This can be anything from limited access to the internet for finding and booking appointments, to language barriers encountered when dispelling myths and controversies around vaccines, to limited ability to travel to vaccine or test sites. A study uncovered which socio-economic factors affected the Covid-19 racial vaccine disparity (CVD) rates affecting 51.5% of the US population, from April 2021, measured as the rate of one vaccine dose for White individuals over Black individuals per US county [**1**]. Vaccine hesitancy was not found to be a significant driver of disparities in vaccine uptake rates within a US county. Rather, the three major social determinants correlated with vaccine disparity were median income, high school graduation rates, and political ideology. They compared with flu vaccine disparity (FVD), and uncovered that the Republican vote share is significantly negatively associated with CVD, much more than for FVD, highlighting the effect of political discourse on social media on vaccination rates around the 2020 election.

It is the responsibility of governments to account for such factors when organizing something as important as distributing life-saving resources. These same social structures, when analyzed on a global scale, are also responsible for the delayed distribution of the vaccine in the global South [**88**]. Clearly, if uptake rates of vaccines are so unfairly distributed, both nationally and globally, it is essential to re-evaluate vaccine distribution systems with a strict focus on access; no community should have less of an opportunity to obtain a vaccine should they choose to do so.

Our proposed methodology therefore has a focus on equitable vaccine access, taking into consideration the demographic distribution of racial groups and income levels in urban areas. Our method selects neighborhoods to be prioritized in the first vaccine roll-out, based on their influence in spreading the disease and fair access considerations. Practically, the use of our method would mean deploying mobile vaccine sites to these neighborhoods, or facilitating their local clinics and hospitals with allocated vaccine budgets. This would also facilitate the community by removing the necessity to travel to obtain a vaccine. Similar work was actually deployed in the state of Virginia, which targeted areas of low vaccine up-take rates [**61**]. Similarly, a mobile vaccination program in Middle Tennessee successfully administered vaccines and information to underserved communities [**3**]. This evidence suggests that our method which uses mobile vaccine sites could similarly be easily deployed.

1.1. Contribution

We firstly propose a solution using a technique called Influence Maximization (IM). Influence Maximization is the optimization problem of finding the most influential nodes in a network - in our case, the most influential neighborhoods in terms of spreading disease in a wider urban area. IM differs from other network science techniques in that it assumes access to a known propagation model used to simulate the spread of the transmissible quantity, for example a disease or information in a social network. The spread function can then be used to run simulations over the network, and spreading occurs due to interactions between nodes in the graph. Such interactions can vary over time, as in our case, where we use real data of visits of neighborhood populations to public places where the disease can spread and infectious individuals can expose susceptible people. We use IM as a basis for our proposed methods, to firstly offer a solution that is competitive in terms of reducing overall infections in the whole urban area, by targetting the most influential communities for vaccination. We subsequently build on this methodology to apply fairness considerations, using Census data to match neighborhoods with demographic data.

1.1.1. Demographic fairness

With the introduction of AI used to inform critical decision-making in high-stake situations, such as loan applications and granting bail to prison inmates based on recidivism risk, many are concerned with the possible social inequalities as a result of the model outputs. Investigative works found the potential of AI systems to exacerbate existing racial and social biases for already marginalized groups [**15, 62**]. ProPublica exposed COMPAS's higher false negative rates for Black inmates when predicting recidivism in the US, meaning more Black candidates were being wrongly predicted of recidivating than White inmates [**6**]. Examples like these have catalysed the research direction of fairness in AI and machine learning, investigating how models can be harmful when using sensitive demographic data, and what methods we can use to mitigate it. In general, fairness methods are categorised into pre-processing (mitigating bias in data), in-processing (mitigating bias in the models themselves), and post-processing (mitigating bias in predictions output from models). In this work, we use an in-processing technique to ensure fairness during the selection of neighborhoods for vaccination. Following previous work in fair influence maximization, we leverage the notion of *equal treatment*, where we ensure fairness in vaccine allocation for social groups, and also measure *equal outcome*, corresponding to fairness in subsequent infection rates for different social groups. Both of these fairness notions exist as pre-defined group fairness notions in the IM literature, but have never been applied to a vaccination task.

Previous research alluded to the idea that economically marginalized communities were less able to reduce their mobility during lockdown [**83, 17, 89**]. Many factors could be responsible for this, in particular how less economically prosperous communities assume more roles of frontline and essential work, within which reducing mobility and working from home is not an option. Within our framework, we target the most influential communities in terms of disease spread. It is therefore possible that racial minority communities would be favoured for vaccination if they are more mobile. In our experiments, though we do find that minorities were less able to reduce their mobility as quickly during lockdown, this does not necessarily lead to their over-sampling when identifying the most influential communities to vaccinate. We posit that this could be attributed to our method reaching as many corners of the network as possible, including for example more remote suburbs of metropolitan areas which are less diverse [**52**].

Many people in non-Western countries also experience vaccine access inequality, for the same reasons that racial minorities in Western countries do; economic and financial disparity is the main cause of boundaries in healthcare access [**66, 42**]. Our framework therefore also focuses on fairness with respect to social groups divided by income level. Indeed, in many societies this will be correlated with racial minorities, as institutional racism historically - and still does - forces migrants into lower economic statuses. However, we also want our framework to be adaptable for countries where there may not be such high racial diversity, or perhaps when statistics on racial diversity by neighborhood are not well documented. We therefore include fairness by economic status, which can be also a useful proxy for fairness by race even for when racial diversity is low [**53**].

Our proposed methods also accommodate for older, higher-risk communities. Many governments opted for vaccinating oldest communities first, since they are more at risk of severe symptoms and death if they contract the disease [**73**]. We therefore build the same prioritization for older groups into our method, using the median age of neighborhoods. In our findings, we uncover how priorization of older generations in the US is sub-optimal in terms of reducing overall infections, due to older groups being less mobile in the network. We also find that only prioritizing age and not mobility can lead to a non-diverse choice of communities selected for vaccination, as older generations in the US are majority White. We therefore propose methods to combine all three aspects; mobility for performance, prioritization of older groups, and fair access for of racial and income groups.

Our main contributions are therefore as follows:

(1) A novel vaccine distribution method using influence maximization to target the most influential neighborhoods, modelling disease spread with a propagation function calibrated to real case counts of Covid-19.
(2) Fairness considerations to achieve equal vaccine access for racial groups and income groups, based on their population size in the network.
(3) Priotization of older neighborhoods to protect those most at risk of severe illness and death, using a novel metric of infections weighted by the median age of the neighborhood.
(4) Experiments on temporal mobility networks constructed from data of real visits occuring during the Covid-19 outbreak in three major metropolitan areas in the US.

Chapter 2

Related work

2.1. Influence Maximization

Influence maximization defines the optimization problem of selecting the k most influential nodes in a graph $G = (V, E)$ from its set of nodes V, in terms of propagating a transmissible substance. Such a substance could be, for example, information or advertising, with the goal of causing the most spreading and maximizing the number of nodes "infected" with the substance. It is distinguishable from other techniques for finding influential nodes since we assume access to a propagation or influence function, σ, which describes how the transmissible substance travels between nodes in the network. The problem solution is a final set of nodes S, often referred to as the *seed set*. This represents the optimum choice of influential nodes, to whom the substance should be given, and who are responsible for transmitting it to the rest of the network. The problem is often applied to viral marketing and resource allocation settings, but is also cited in epidemiology literature. In our application, we use IM to identify the neighborhoods in an urban area who cause the most infections in the whole network, when simulating Covid-19 beginning from each neighborhood at a time.

Solving an influence maximization problem exactly and providing optimal solutions has been shown to be NP-hard [**44, 55**], and therefore many approximation algorithms have been proposed to reduce the complexity of the problem. Most notably, Kempe et al. [**44**] demonstrated how the greedy IM strategy provides a $1 - 1/e$ approximation guarantee, so long as the influence function is both monotone and submodular. That is, for a greedy solution $\hat{S}$, and true optimal solution S^*, we have $\sigma(\hat{S}) \geq (1-1/e)\sigma(S^*)$. The classic greedy algorithm is shown in Algorithm 1.

The greedy approach can also be adapted to the scenario where each node u is attributed with its own cost, $c(u)$. In this case, we want to maximize the influence in the network caused by selected nodes S, but are bound by a budget B, requiring $\sum c(u) \leq B, \forall u \in S$. In this work, we are interested in this application since we are bound by a total vaccine

Algorithm 1 Greedy influence maximization [**44**]

Input: $G = (V,E)$, k, σ
Output: S
$S \leftarrow \emptyset$
for $i = 1,...,k$ **do**
    $u^* \leftarrow \text{argmax}_{u \in V \setminus S} \sigma(S \cup \{u\}) - \sigma(S)$
    $S \leftarrow S \cup \{u^*\}$
end for
return S

budget for the entire network, and each neighborhood has its own "cost" - its population size. Allocating vaccines for a higher population neighborhood is therefore more costly to the vaccine budget than a lower population neighborhood. Subsequent works including CELF [**50**] and CELF++ [**32**] focus on this cost-adapted setting, but also offer more efficient implementations of the greedy strategy. In Section 3.1, we outline how CELF saves on compute time by exploiting the submodularity of the influence function. Like these, there are many following works which attempt to reduce the hardness of the IM problem, and can generally be categorized into simulation-based (running Monte Carlo simulations) [**91, 84, 38**], proxy-based (devising proxy propagation models for σ) [**20, 67**] and sketch-based approaches [**51, 22**].

2.1.1. Submodularity

Many works which choose to adopt a greedy strategy must therefore rely on proving the submodularity and monotonicity of their influence functions, in order to provide the same optimality guarantees outlined by Kempe et al. In typical static graph settings where nodes and edges do not change over time, proofs of submodularity are typically established by leveraging the notion of reachability [**55, 56**]. For example, it is possible to generate multiple graph instances which illustrate the "path of infection". Certain nodes may begin with the disease at $t = 0$, and at each time instance there is some probability of infecting its neighbors with a certain probability p. We can therefore build a subgraph instance where each edge $e(u,v)$ is placed in the graph if, when sampling from probability p, node u infects node v. The resulting graph is therefore a mapping of the infection path from $t = 0$ to the final time instance $t = T$. Any nodes with which there is a defined path to the initial infected nodes are therefore "reachable" by the infected nodes. Taking multiple of these subgraph instances and averaging counts of reachable nodes is a way of estimating the influence of a set of seed nodes S. While this aids submodularity proofs for typical static graphs, this approach is not applicable to our setting, where one edge represents the number of people moving from a neighborhood to a public place. It is therefore not possible to track the

infection paths, as the granularity of the data does not allow us to track movements of specific infected individuals, only the movement of fractions of the neighborhood together.

Gayraud et al. [**31**] pose an interesting problem with establishing submodularity on diffusion functions on temporal networks, in which edges change over time. They demonstrate that the total infections caused in the network will depend on the placement of the starting infected nodes. In some cases, because edges are added and removed, some parts of the graph will be infectious but be "locked-off" from the rest of the network. The infectiousness of the infected nodes may expire during this time, and by the time edges appear to connect these nodes to the rest of the network, the infectiousness has already worn off. This results in a trapping effect of the infection, which means that the placement of the initial starting nodes affects how many infections can occur. They demonstrate with examples that this causes the propagation function to be neither monotone nor submodular in some cases.

Contrary to this, Erkol et al. [**25**] demonstrate that the greedy strategy can still be effective even when the influence function is not always submodular, but the number of submodularity violations is low. In our work, we also observe low submodularity violations, and therefore maintain using the greedy strategy.

2.1.2. Fair influence maximization

Many works have proposed solutions for introducing demographic fairness to influence maximization problems. This problem is relevant to settings where each node u has its own demographic attributes. In many cases, nodes can represent individuals, for example in a social network; we may then have access to their demographic information, such as race, gender or age. The specific fairness task will depend on the choice of fairness definition in the IM setting. For example, *equal treatment* tries to achieve proportional sizes of each social group within the seed set S. *Equal outcome*, on the other hand, requires proportional sizes of each social group within all the nodes who are infected. The choice of fairness definition will depend on the application, and whether nodes in the seed set S are particularly benefited. In our case, our seed set nodes experience a benefit because they are selected for vaccination first; we therefore employ equal treatment for racial groups and income groups in the network. However, we also measure the equal outcome in infections as a result of our vaccine strategy.

Farnadi et al. [**28**] present a framework for solving the fair IM problems optimally with mixed integer programming. Their framework allows for an array of fairness constraints in IM, including *equal treatment* and *equal outcome*. Though the solver finds optimal solutions, the method is not scalable for large networks. Anwar et al. [**7**] present a methodology for ensuring fairness in networks where nodes can belong to two or more defined groups, and display evidence of homophily in their interactions, i.e. members of the same group are more

likely to interact with each other. In this setting, they demonstrate with use of a fairness regularizer on the influence function that it is possible to ensure fairness in terms of *equal outcome*. This constraint aims to ensure that the number of influenced nodes in a group is proportional to that groups size in the entire network.

Within the IM literature scope, the closest work to our contribution is from Minutlo et al. [**65**], who lay the theoretical foundations for using influence maximization for targeted vaccination. Our contribution differs significantly from this, as theirs does not consider demographic fairness, whilst our contribution is motivation by fair vaccine access. Their work does not model any specific disease as we do, and in fact is simplistic in using generic IM propagation functions on networks unrelated to any specific disease. They also experiment on networks with interactions between individuals, whereas our community-level structure is a more condensed way to model the interactions between millions of people with only a few thousand neighborhood nodes. Further, our network setup is more realistic for obtaining demographic information; privacy concerns rightly limit the possibility of reconstructing both demographic and mobility information for individuals, but is available publicly through Census data at the neighborhood level.

2.2. Vaccine distribution optimization

Many works have discussed how to optimize vaccine distribution in a setting with limited supply by use of mathematical modelling commonly used in epidemiology, particularly around the period of the Covid-19 vaccine first becoming available [**57, 60, 11, 10**]. Matrajt et al. [**59**] find that in the situation of low vaccine effectiveness, it is most beneficial to vaccinate older groups first to minimize overall deaths. But in the setting of high vaccine effectiveness, it is more effective to vaccinate high-transmitting individuals, which generally aligns with younger populations. Similarly, Shim et al. [**81**] and other works find that the most effective age group to vaccinate depends on the objective; typically younger adults, around 20-49 years, to minimize infections and those above 50 to minimize mortality rates. The importance of time-varying prioritization strategies is also stressed, in order achieve multiple objectives like minimizing overall deaths and infections simultaneously [**34**]. In our work, we also consider multiple objectives to optimize for, by creating metrics that combine minimizing overall infections and protecting the most at-risk (older) groups.

There are also works that consider targeted immunization of the most influential disease spreaders to optimize the vaccine strategy [**35**]. Lee et al. consider time-varying interactions in a network [**49**], and detects the most influential members of a network by looking at historical contact patterns and projecting into the future. However, this requires precise naming of interactions between individuals, which are difficult data to obtain in the real-world. In our

work, we combine these efforts to provide a solution that both considers the most influential communities, protects the most at-risk communities *and* considers demographic fairness.

Mehrab et al. [**61**] propose a method for sending mobile vaccine sites to areas with low up-take rates and high hesitancy, and use demographic information to support minority racial groups. Their solution is more of a post-mortem approach for supplementing vaccines in resulting gaps in up-take rates after deploying the initial round of vaccines. Arguably, the delay in time between the first distribution of vaccines and the deployment of their post-mortem mobile vaccine sites allows for infections to spread more amongst the minority groups and inequality to occur. Our proposal however aims to ensure fair vaccine access for all social groups during the first rounds of roll-out from the beginning. Their method was actually deployed in practise to assist the Virginia Department of Health in 2021, which strengthens the plausibility of deploying our method in the real-world, since this would also rely on mobile vaccination sites to target specific neighborhoods.

The work by Chang et al. [**17**] simulates Covid-19 spread on mobility networks to measure the effects of the lockdown strategies, including tests of counterfactual lockdown implementations. Their study focuses on the importance of lockdown for various categories of points of interest, such as restaurants and gyms. However, the results demonstrate how racial minorities were less able to reduce their mobility during lockdown, resulting in unequal rates of infection depending on race. These findings largely motivated the work of our paper, to consider how different demographic groups may exhibit different mobility patterns, and whether a vaccine solution can cater to this. In our work, we leverage some contributions made by Chang et al. [**17**], using their method to construct mobility networks using real visit data in the US, and simulate their Covid-19 model. We adapt this framework to our setting by introducing vaccination into the model, and including age information per neighborhood before conducting our fair influence maximization method.

Chapter 3

Additional Preliminaries

In this chapter, we outline some additional theoretical and practical details which motivate our approach. The reader may find it useful to first read Section 4.3 of the main article to familiarize themselves with the problem setup before continuing with this chapter.

We firstly present some justificiation of our choice of the greedy approach with Monte Carlo simulations over others in the literature. The existing literature on IM for temporal networks and similar diffusion models predominantly focuses on the greedy approach [**87**]. Other methods that utilize diffusion model proxies are primarily designed for simpler diffusion functions like the Independent Cascade and Linear Threshold models [**79, 21**], typically seen in IM literature, and do not readily extend to more complex models. Our SEIR (susceptible, exposed, infected and recovered) model, as outlined in the Appendix (Section 4.8), is notably complex and features non-linear dependencies on the matrix of visits data. Consequently, attempting to reliably simplify this function would prove infeasible. Regarding alternative sketch-based methods, our mobility network presents a unique challenge. We cannot meaningfully construct sketches (graph instances) since our edges represent the movements of multiple individuals conducting visits, not just a single person. As a result, it becomes impossible to track the definitive paths of infection created by infectious individuals for the purpose of evaluating influence, given that their movements cannot be isolated. For these reasons, we opt for the greedy approach with Monte Carlo simulations, and we present further implementation details below.

3.1. CELF

Classically, in order to provide optimality guarantees, the greedy IM strategy imposes two constraints on the influence function - monotonicity and submodularity. The contribution made by Leskovec et al. [**50**] was to leverage the submodularity of the influence function in order to reduce the number of calculations required to execute the greedy IM

strategy. We also implement this contribution, named Cost-Effective Lazy Forward (CELF), and here explain why it reduces computation time for IM with submodular functions.

Monotonicity: A set function f is monotonic if for every $X \subseteq Y$ we have that $f(X) \leq f(Y)$.

This requires that the function f when applied to a set X is never greater than the output of the same function when applied to a superset of X, Y.

Submodularity: A set function f is said to be submodular if for every $X, Y \subseteq V$ with $X \subseteq Y$ and every $x \in V \setminus Y$ we have that $f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$.

In words, the gain in function f when adding a new set element x to set X is greater than the gain in f when adding x to set Y, if X is a subset of Y. This is often referred to as the law of diminishing returns; the gain to be achieved is decreasing as we add to the size of the set. We refer to the "marginal gain" of adding the set element x to set X as $g^X(x) = f(X \cup \{x\}) - f(X)$.

During the greedy implementation, recall that one new node is added to the seed set at a time, because that node provides the greatest marginal gain compared to all the other candidates at that iteration (see Algorithm 1). The size of the seed set is therefore growing with each iteration. In this case, if our influence function is submodular, the possible gain of adding a new node u_i, if it has not already been added, is always diminishing as the seed set grows in size: $f(X \cup \{u_i\}) - f(X) \geq f(Y \cup \{u_i\}) - f(Y)$. The trick that Leskovec et al. realized is as follows: suppose that a node u_i is added to the seed set S at iteration n, where $n > 1$, because it provided the greatest marginal gain at that iteration. Then, at iteration $n + 1$, the seed set is now $S \cup \{u_i\}$ after adding u_i in the previous round. Also suppose the node with the second greatest marginal gain in iteration n was node u_j. Then, if only u_j's gain is re-calculated in round $n + 1$, with the modified seed set $S \cup \{u_i\}$, and it is the highest gain node compared to all the other nodes whose last gain calculation is with a smaller version of S, then it is in fact the highest gain node of iteration $n + 1$ compared to all the other candidate nodes, and we do not need to re-calculate the gain of all the other nodes to prove it.

Why is this true? Consider the node with third greatest gain at iteration n, u_k - see the diagram in Figure 3.1. We use S_n to denote the set S at round n, so S_1 denotes S at the first iteration. The gain $g^{S_1}(u)$ therefore denotes the gain of node u evaluated when S was in the first iteration. Some nodes will not have had a more recent gain calculation since the first round, so we illustrate those here. As the seed set grows from iteration n to $n + 1$, the gain of adding u_k to the seed set can only decrease, due to submodularity (which is true of all

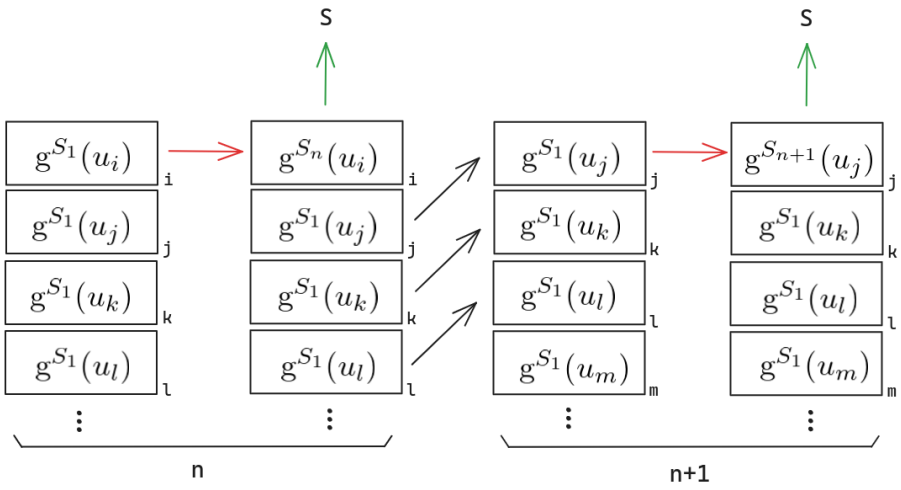

Fig. 3.1. A diagram of the greedy node selection using CELF [**50**]. Nodes (rectangles) are ordered by their marginal gain on the seed set g^S. At each iteration, the node with the greatest gain from the last iteration (n) has their gain re-calculated for the current iteration ($n+1$), which may change its position in the list (red arrow). If it maintains its top position, it is added to the seed set (green arrow).

nodes). In this case, it can never supersede it's gain from the previous iteration n. Hence if u_j's gain at round $n+1$ is greater than u_k's gain at any other previous round, it is definitely greater than u_k's gain at round $n+1$. Mathematically, due to submodularity, we know that the gain of any node reduces as the seed set grows, i.e. with iterations. We can write this as

$$g^{S_1}(u_k) \geq g^{S_{n+1}}(u_k), \tag{3.1.1}$$

for node u_k, though it applies to all nodes. And if, when evaluated at round $n+1$, when S is larger, the gain of u_j is greater than the last evaluation of u_k's gain, then we can write

$$g^{S_{n+1}}(u_j) \geq g^{S_1}(u_k) \geq g^{S_{n+1}}(u_k). \tag{3.1.2}$$

We can therefore infer that u_j is indeed the highest gain node at round $n+1$, without having to re-calculate the gain of u_k or any of the other nodes in the list. This saves a significant amount of time and compute power, as we are not required to run simulations to calculate the new gain value for all nodes at every iteration of IM.

It's possible that, for example, after re-calculating node u_j's gain at iteration $n+1$, is not greater than u_k's last evaluated gain from a previous round. This indicates that u_j is not the optimum node choice at $n+1$, since the influence of some other nodes calculated in a previous round have superseded it. In which case, we re-calculate whichever node appears at the highest gain position in the list until it maintains the highest position spot after calculating their gain for the $n+1$th iteration.

In our application of this method, the disease spread function is generally submodular, but some violations are exhibited in a few rare cases. We therefore implement CELF, and the corresponding re-calculation steps can be seen in lines 13 to 18 of Algorithm 2 in the main article. However, when we later amend our spread function to protect older populations (see Section 4.4), we witness many more submodularity violations, and therefore remove the re-calculation steps.

3.2. Vaccination model

In this Section, we present the intuition for how the disease propagates according to the model proposed by Chang et al. [**17**], and how we adapt this to model vaccination in the network. Figure 3.2 depicts the bipartite graph structure used to represent the mobility network. Edges connect a CBG (Census Block Group - a geographical divisions of a neighborhood in the US) to a POI (Points of Interest - public places such as gyms, restaurants etc.), with a weight corresponding to the number of people from the CBG visiting the POI at hour t. The edge weights therefore vary hourly, with the possibility of reducing to zero, which indicates no visitors at the POI were from that particular CBG during that hour. The disease is able to spread due to infectious individuals mixing with susceptible individuals at POIs, depending on the average duration of stay and the number of infectious people at the location. It is also possible for individuals to be infected when they are not visiting POIs, which can represent for example infections at home or on public transport - anywhere that is unaccounted for in the potential public places. The full disease model equations are included in the Appendix of the main article, see Section 4.8.

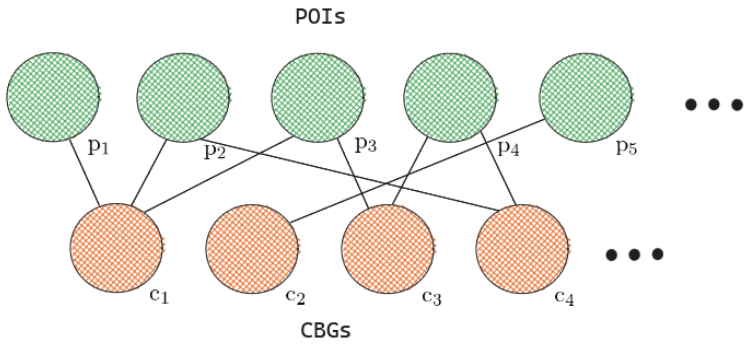

Fig. 3.2. Bipartite graph of CBG and POI nodes. Edge weights e_{ij}^t represent the number of individuals from CBG c_i visiting POI p_j at hour t.

3.2.1. Parameter learning for propagation model

We initialize the SEIR model following the work of Chang et al. [**17**], with the full mathematical model outlined in Section 4.8. To ensure the model exhibits realistic spread dynamics, it is calibrated to match real case counts of Covid-19 over the same time period in each of the different MSAs (metropolitan statistical areas). The SEIR model contains several variables which affect the likelihood of being infected, such as the number of infectious visitors at a POI. However, there are only three free parameters which are used to fine-tune the model's predicted number of infections with real case counts per each metropolitan area. These free parameters are

(1) p_0, the probability of any individual being infected at the first hour t_0 of the simulation,

(2) ψ, a transmission constant shared across all POIs, which affects the likelihood of being infected at a POI, and

(3) β_{base}, the base transmission rate shared across all CBGs, which affects the likelihood of being infected while not at a POI, for example while at home or on public transport.

After finding plausible ranges for these parameters, a grid search is performed to fit the output of the model to real case counts.

The confirmed case counts are only used to reflect 10% of all case counts, including those unreported. Therefore, only 10% of total simulated infections are used to match to real cases. The predicted case counts are also modelled to be confirmed 7 days after initial infection. For each MSA, we select the set of calibrated model parameters with the lowest RMSE with real case counts, summed over each day of the simulation. The final model parameters for each MSA are included in Section 4.8.

3.2.2. Simulating vaccination

We will now describe how we manipulate the simulations to implement IM and evaluate our results. Figure 3.3 shows how disease spread amongst CBGs is modelled in the network. For the sake of the demonstration, we have simplified the bipartite graph of CBGs and POIs in Figure 3.2 to only containing CBG nodes, and creating an edge at hour t between two CBG nodes if members from both those CBGs are present at the same POI at that hour. In this case, there is a possibility of infection spread between members of the two CBGs. The diagram demonstrates a toy example of the disease spread mechanism from $t = 0$ to $t = 2$, on a network of six CBGs. At each time step, the edges and their weights change, in some cases disappearing entirely (edge weight of zero). This edge weight signifies the number of people overlapping at the same POI at hour t between two CBGs. An edge weight of zero therefore indicates no members overlapping at the same public place.

In order to quantify the influence of each of CBGs in terms of its capacity to spread disease, our proposed algorithm runs what we define as "prior" simulations. Figure 3.3 (a) demonstrates how the prior simulation works, where we model disease spread when members of only one CBG have the possibility of being infected at time $t = 0$. This probability is the parameter p_0, the probability of infection at $t = 0$ - one of the three model hyperparameters which is tuned to ensure the simulation matches real case counts when setting up the mobility networks. For the other remaining CBGs in (a), p_0 is set to zero. We run this simulation n_{seeds} times for every CBG, taking an average over the resulting infections at the final time T. The result is an averaged number of infections caused due to a particular CBG beginning with the disease. This is how we first create an ordered list of the CBGs by their "prior" influence, before moving onto the stage where we iteratively select from this list which CBGs are most *jointly* influential and therefore should be vaccinated.

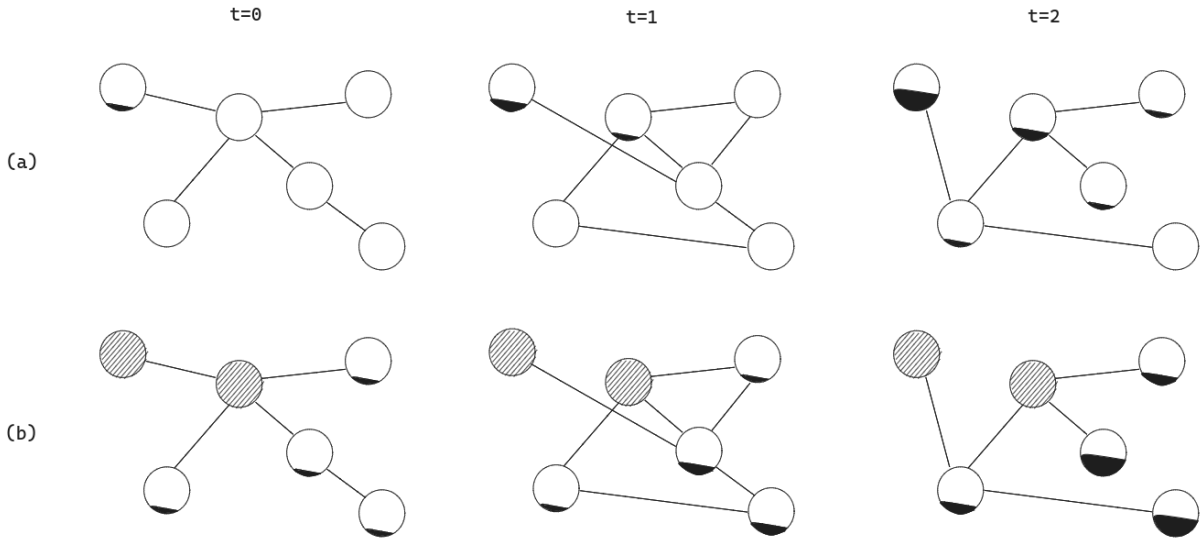

Fig. 3.3. Spread of disease exposure amongst CBGs in (a) prior simulations, where individuals in only one CBG have a small probability of being infected at $t = 0$, and (b) evaluation simulations, where members of all CBGs have a small probability of infection at $t = 0$, except those that are vaccinated. The dashed CBGs represent those that are vaccinated, and the black filled areas represent the fraction of the CBG members who have been exposed to the disease (the size is exaggerated for these early time steps to illustrate the point).

Once we begin iteratively adding CBGs to the set Z selected for vaccination, as mentioned in Section 3.1, more influence calculations are required as the set Z grows in size. Here, we use Z to denote the seed set instead of S as is typical in the IM literature, because in our case we use S to denote the "susceptible" state in the SEIR model. In this case, we are calculating the gain of adding a particular CBG c_i to a selected set Z, given by $g^Z(c_i) = \sigma(Z \cup \{c_i\}) - \sigma(Z)$. To calculate the influence of multiple CBGs, we therefore run prior simulations, but with all the CBGs in Z starting with infected individuals at $t = 0$, instead of just one CBG. This

would look like Figure 3.3 (a), but with multiple nodes starting with the disease instead of just one. Then, the probability of infection for individuals in CBG c_i at time $t = 0$ is

$$p_0^{c_i} = \begin{cases} p_0, & \text{if } c_i \in Z \\ 0, & \text{otherwise.} \end{cases} \tag{3.2.1}$$

Figure 3.3 (b) illustrates a different simulation type, which is used during the evaluation stage. In the evaluation phase, a set of CBGs V have been selected for vaccination (V denotes the final version of Z), depicted as dashed nodes in (b). We want to measure the total infections in the network at the final time T as a result of vaccinating set V. In this case, individuals in all the remaining, non-vaccinated CBGs can begin infected at time $t = 0$ with probability p_0, such that

$$p_0^{c_i} = \begin{cases} p_0, & \text{if } c_i \notin V \\ 0, & \text{otherwise.} \end{cases} \tag{3.2.2}$$

The disease can spread accordingly using the hourly visits, with vaccinated CBGs never being able to contract the disease. We ensure that vaccinated CBGs never contract the disease by setting the number of susceptible individuals to zero at all time periods in those CBGs. We also take averages of these results with $n_{seeds} = 10$.

An alternative approach to this optimization problem, which seems more relevant to the problem, is to instead employ an influence function to maximize the *prevented* infections due to vaccination, rather than selecting CBGs which cause the most infections and then vaccinate them. In our experiments, explored utilizing an influence function designed to maximize infections prevented by vaccination. However, the inherent high uncertainty associated with our diffusion model necessitates a significant increase in the number of Monte Carlo simulations required to accurately estimate the impact of vaccination, as compared to estimating infections caused. We found that measuring the infections prevented by vaccinating only one CBG at a time was not tractable without many more random runs. This substantial computational overhead renders the simulation unnecessarily inefficient and resource-intensive.

Additionally, the limited knowledge available surrounding the performance of the vaccine in the early stages of pandemics would introduce further uncertainty into our modelling. Variables such as vaccine effectiveness rates and the time to achieve immunity may not be as extensively studied before the first round of large-scale deployment. Consequently, if we were to base our CBG selection on infections prevented by vaccination, we would exacerbate the existing uncertainties within the model. This would compromise the realism and practicality of our model, making it less suitable for its intended purpose.

These considerations underscore the challenges and limitations associated with adapting the influence function to maximize the infections prevented by vaccination. For these reasons, we believe that our current methodology remains more suitable for this application.

3.3. Replacing social distancing data

In the original framework by Chang et al. [**17**], they use social distancing data provided by Safegraph in 2020 which contained foot traffic data during the pandemic. One portion of this data contained the fraction of each CBG who stayed at home daily, which was used in order to construct the mobility networks. This particular section of the data was discontinued by Safegraph, so instead we reconstruct the same stay-at-home data using the Neighborhood Patterns data which Safegraph still provides.

The Neighborhood Patterns documentation, found here [**78**], provides explanations for features in the data. In this section we will demonstrate how we extracted the daily counts of CBG members who were out visiting other CBGs i.e. those who did not stay at home per day. In the Neighborhood Patterns dataset, there is information about stops made by individuals and unique devices stopping at a particular CBG, including which home CBG the visitors are from. The intuition is to aggregate over this to extract a count of how many individual devices per home CBG were out visiting other CBGs everyday.

One device corresponds to a unique individual, but each device can make multiple stops in the same CBG. So firstly, we calculate a factor g per CBG, defined as the ratio of stops made to devices stopping at that CBG, an indication of the average number of stops people make at a CBG. We have access to `raw_stop_counts` and `raw_device_counts`, which provide the number of stops and devices respectively *visiting* a CBG during the total time period (5 weeks). We use this to calculate an average number of stops per device in that area, which we call the factor g_i,

$$g_i = \frac{\texttt{raw\_stop\_counts}_i}{\texttt{raw\_device\_counts}_i} \tag{3.3.1}$$

which is defined for each CBG c_i. We use this factor to then convert `stops_by_day`, the count of stops in the area per day, to a daily device count d_n for the nth day in our total time period, given by

$$d_i^n = \frac{\texttt{stops\_by\_day}_i^n}{g_i}. \tag{3.3.2}$$

This gives us, for each CBG c_i, a count of the number of people (devices) stopping there for each day n in our total time period.

The `weekday_device_home_areas` and `weekend_device_home_areas` contain counters of the number of devices stopping in a CBG grouped by their home CBG, during weekdays and weekends respectively. We turn these into distributions $D_i^{weekday}$ and $D_i^{weekend}$ of the

number of devices stopping at CBG c_i coming from other home CBGs. Sampling once from these distributions will therefore return a device labelled with their home CBG, based on the probability that members of that CBG visit CBG c_i on weekdays or weekends respectively. We choose to separate the weekdays and weekends since visit patterns change based on whether people are commuting for work or not.

For each of the n days in the total time period, we then sample a total of d_i^n daily devices from the distributions $D_i^{weekday}$ or $D_i^{weekend}$ depending on which day of the week day n is. We then have daily CBG samples s_i^n, a vector of size d_i^n, where each element represents one device from a particular home CBG visiting CBG c_i on day n. In vector s_i^n, there will be several elements containing the same home CBG, representing different visitors (devices) from the same CBG. We can then turn s_i^n into an un-normalized distribution of device counts visiting c_i, S_i^n, for all n days.

For each of the n days, we can then use all of these distributions S_i^n to count how many visits are *made* by devices from a home CBG c_j, to any other CBG (switching the index to j to signify these are *home* CBGs, not CBGs being visited). This provides a count, per home CBG c_j, of how many devices were out visiting any other CBG c_i on each day, $\texttt{device\_count\_out}_j^n$. We can therefore compute the fraction of a CBG which did not stay home, `fraction_out`, as

$$\texttt{fraction\_out}_j^n = \frac{\texttt{device\_count\_out}_j^n}{n_j} \tag{3.3.3}$$

where n_j is the total population size of CBG c_j. The fraction of CBG c_j which stayed at home on day n is therefore $1 - \texttt{fraction\_out}_j^n$. We implement this in place of the discontinued social distancing data used in the previous work, an essential element to constructing the mobility networks accurately.

3.4. Multiple Knapsack Problem

In Section 4.4, we outline how we distribute vaccines amongst social groups to achieve fairness in terms of equal treatment. We formulate this task similarly to the Multiple Knapsack Problem, with some amendments, which we detail below.

In general, the multiple knapsack refers to the optimization problem in which we want to maximize the profit contained in M knapsacks, by filling them with objects, choosing from K potential items with profits p_j and weights w_j. We are constrained by not exceeding the maximum weights W_i of the m knapsacks. The factor x_{ij} refers to the fraction of item j which we choose to store in knapsack i.

In our work, we split the total vaccine budget B for the whole network into separate vaccine budgets for each social group. We do this by taking proportional numbers of vaccines

per social group, according to the fraction of that social group in the entire network, such that $B_j = \frac{N_j}{N} B$, where N_j is the population of social group j in the whole network, and N is the whole network population. The multiple knapsack is therefore analogous to our problem, since we are trying to distribute vaccines up to the vaccine budget W_j (B_j), for each of the M social groups, by selecting from K CBGs (items) which add the most influence p_i (profit). Each CBG contains a certain population (weight) w_i, which is comprised of populations of the M social groups, i.e. each CBG has its own demographic makeup of racial groups; x_{ij} therefore denotes the fraction of CBG c_i's population belonging to social group j. When a CBG is selected for vaccination, we can think of filling the knapsacks by separating the CBG population into it's M social groups, and filling each corresponding social group "knapsack" accordingly.

We make some small adaptations to the optimization problem, in that each population fraction x_{ij} can exist in the range 0 to 1, for example a CBG has 40% Black/African-American residents, so the x_{ij} is 0.4. The value of these factors are therefore not modifiable. Further, each CBG (item) can either be selected for vaccination or not. In this sense, the sum of its fractions x_{ij} over the social group knapsacks is either 0 or 1, a binary indicator for whether it has been selected. The factors are therefore modifiaible only in the sense that they can be zero (CBG not selected) or the constant fraction of social group j in CBG i. The final optimization problem is therefore given by

$$
\begin{aligned}
\text{maximize} \quad & \sum_{i=1}^{K} \sum_{j=1}^{M} p_i x_{ij} \\
\text{subject to} \quad & \sum_{i=1}^{K} w_i x_{ij} \leq W_j, \quad \text{for all } 1 \leq j \leq M \\
& \sum_{j=1}^{M} x_{ij} \in \{0,1\}, \quad \text{for all } 1 \leq j \leq K \\
& x_{ij} \in [0,1] \qquad \text{for all } 1 \leq j \leq K \text{ and all } 1 \leq i \leq M
\end{aligned}
$$

In the implementation, our greedy IM approach already takes care of ordering the CBGs by maximium gain (profit) at each iteration. The only change we need to make is therefore simply filtering out the candidate CBGs which would exceed any of the social group budgets W_j if they were to be chosen for vaccination. This is implemented at the end of each IM iteration, between lines 19 and 20 of Algorithm 2. The result is therefore only selecting from the most influential candidates at each iteration which do not exceed any of the social group budgets.

Chapter 4

Promoting Fair Vaccination Strategies through Influence Maximization: A Case Study on COVID-19 Spread

by

Nicola Neophytou[1], Afaf Taïk[1], and Golnoosh Farnadi[2]

([1]) Mila, Quebec AI Institute and Université de Montréal

([2]) Mila, Quebec AI Institute and McGill University

This article was submitted in a conference and has passed the first round of review, but still under review for the second round at the time of submitting this thesis.

The main contributions of Nicola Neophytou for this articles are presented.

- Development of ideas and underlying theory with the guidance of the supervisor;
- Design of the open-source framework and implementation of all experiments and simulations on real-world mobility datasets;
- Writing the presented article as first author;

Co-author Afaf Taïk provided useful feedback on the project direction. Co-author Golnoosh Farnadi has guided the research direction as the principal investigator (PI).

Résumé. Les conséquences de la pandémie de Covid-19 ont été plus graves pour les groupes minoritaires raciaux et les communautés économiquement défavorisées. De telles disparités peuvent s'expliquer par plusieurs facteurs, notamment l'inégalité d'accès aux soins de santé, ainsi que l'incapacité des groupes à faible revenu à réduire leur mobilité en raison de leurs obligations professionnelles ou sociales. De plus, les personnes âgées se sont révélées plus susceptibles de présenter des symptômes graves, en grande partie pour des raisons de santé liées à l'âge. Il est donc essentiel d'adapter les stratégies de distribution des vaccins pour tenir compte d'un éventail de données démographiques pour remédier à ces disparités. Dans cette étude, nous proposons une nouvelle approche qui utilise la maximisation de l'influence (IM) sur les réseaux de mobilité pour développer des stratégies de vaccination intégrant l'équité démographique. En prenant en compte des facteurs tels que la race, le statut social, l'âge et les facteurs de risque associés, nous visons à optimiser la distribution des vaccins afin d'obtenir diverses définitions d'équité pour un ou plusieurs attributs protégés à la fois. Grâce à des expériences approfondies menées sur la propagation du Covid-19 dans trois grandes zones métropolitaines des États-Unis, nous démontrons l'efficacité de l'approche proposée pour réduire la transmission de la maladie et promouvoir l'équité dans la distribution des vaccins.
Mots clés : Maximisation de l'influence, réseaux de mobilité, distribution de vaccins, équité démographique

Abstract. The aftermath of the Covid-19 pandemic saw more severe outcomes for racial minority groups and economically-deprived communities. Such disparities can be explained by several factors, including unequal access to healthcare, as well as the inability of low income groups to reduce their mobility due to work or social obligations. Moreover, senior citizens were found to be more susceptible to severe symptoms, largely due to age-related health reasons. Adapting vaccine distribution strategies to consider a range of demographics is therefore essential to address these disparities. In this study, we propose a novel approach that utilizes influence maximization (IM) on mobility networks to develop vaccination strategies which incorporate demographic fairness. By considering factors such as race, social status, age, and associated risk factors, we aim to optimize vaccine distribution to achieve various fairness definitions for one or more protected attributes at a time. Through extensive experiments conducted on Covid-19 spread in three major metropolitan areas across the United States, we demonstrate the effectiveness of our proposed approach in reducing disease transmission and promoting fairness in vaccination distribution.
Keywords: Influence Maximization, mobility networks, vaccine distribution, demographic fairness

4.1. Introduction

The fallout of Covid-19 revealed the stark inequalities in access to healthcare between social groups in diverse and urban areas [**70, 75, 64, 40, 8, 36, 30**]. Studies confirmed that economically-deprived communities and racial minorities experienced higher rates of

infection, hospitalization and mortality as a result of Covid-19 [**47, 82, 2**]. The reasons for this disparity form a long chain of events, with unequal access to healthcare between socioeconomic groups, and therefore racial groups, at the root of it. Furthermore, studies on mobility networks in the US also revealed how minority communities were less able to reduce their mobility as quickly during the pandemic, and as a result suffered higher rates of infection [**17**]. This can be largely attributed to underprivileged groups assuming the roles of frontline and critical infrastructure work, and also living and working in more crowded circumstances. Inequalities in access to the internet and ease of travelling to test and vaccination sites are also factors contributing to this discrepancy [**90**]. These data demonstrate inequities in receiving resources throughout the pandemic, which also extends to vaccination [**74, 43, 69, 9**]. In Figure 4.1, we demonstrate how racial minorities and lower income communities in three US metropolitan areas were less able to reduce their mobility as quickly when the lockdown was introduced.

This disparity motivates the need for a fair vaccination strategy that differs from the current technique. In this work, we investigate a collection of alternative vaccination strategies that consider both mobility and fairness. We leverage an approach called influence maximization (IM), a network science technique designed to detect the most influential members of social networks, typically used in applications such as viral marketing campaigns [**19**]. We adapt this principle to instead detect the neighborhoods or communities which exhibit the largest influence on a mobility network in terms of disease propagation. Such communities are likely to include essential workers who are less able to reduce their mobility during lockdowns [**76, 72**].

We adapt our IM approach to achieve fairness in vaccine allocation for racial groups as well as groups of different social statuses. Moreover, older individuals may be less mobile but more at risk of severe outcomes when exposed to the disease. It is essential not to overlook this trade-off; we therefore also design a strategy designed to protect communities based on higher risk and vulnerability. To summarize, our contributions are as follows:

(1) A novel community-level influence maximization approach for identifying impactful neighborhoods, aiding targeted vaccination against disease transmission.

(2) Extension of influence maximization to mitigate infection disparities among racial and income-level communities.

(3) Introduction of a competitive method, merging influence maximization with prioritizing older communities to reduce overall infections.

(4) Empirical validation on mobility networks from three major US metropolitan areas, utilizing real aggregated visit data from census block groups (CBGs) to points of interest (POIs) during the first five weeks of Covid-19 pandemic.

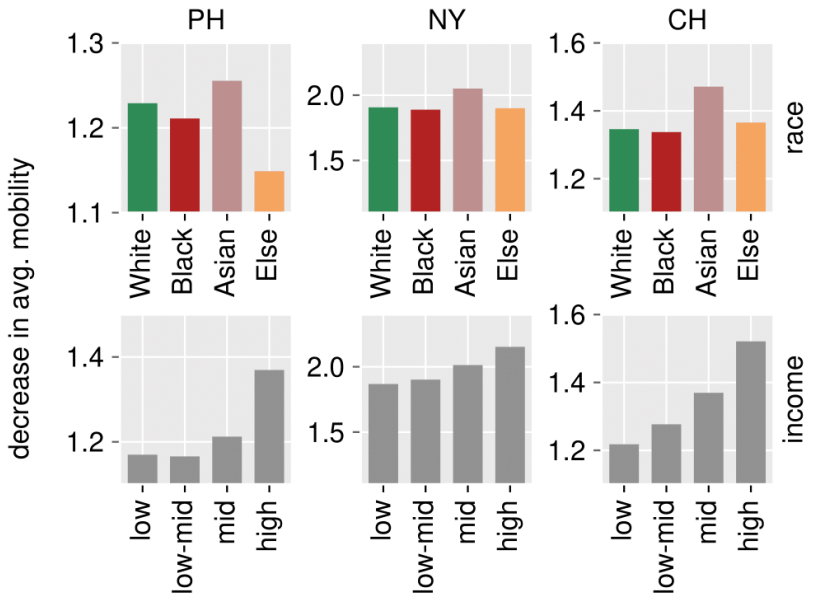

Fig. 4.1. The *reduction* in mobility from before lockdown to during lockdown per racial groups and income groups in Philadelphia, New York and Chicago metropolitan areas. For all three areas, lower income groups and racial minorities belonging to lower income groups (see Fig. 4.2) were less able to reduce their mobility as quickly when transitioning to lockdown.

4.2. Related Work

Fairness in AI is an expanding area of research which has seen traction since the exposure of biases in several significant technologies [**62, 15, 6**]. This extends to the field of influence maximization, where various works impose fairness constraints on the optimization problem. For instance, Farnadi et al. provide a framework for applying a variety of fairness definitions to IM tasks [**28**]. Other works have included developing adversarial graph embeddings to achieve fair IM in social networks [**45**], and balancing majority and minority groups in IM when networks and diffusion processes demonstrate homophily [**7**]. Ali et al. ensure fairness in exposure under a time-critical perspective, for example posting a job advertisement which should be reached by equal sub-populations before the deadline to apply for the position [**4**]. This does not however consider the possibility of changing network structure over time, as is the case with mobility or contact networks used in epidemiology, and is therefore not extendable to the problem of vaccine distribution.

Meister et al. characterize communities by their social activity and vulnerability due to age in a comparative game for optimizing vaccination, however, they do not use real contact or mobility data to test their approach [**63**]. Similarly, works which do consider demographic fairness in vaccination do not necessarily consider mobility to improve performance [**46, 68, 41**]. Anahideh et al. propose a vaccine allocation solution which tackles the apparent

trade-off between equal distribution amongst regions and demographic fairness [**5**]. Their approach defines effective distributions as those that cover many geographical areas; however, we argue that this does not necessarily reduce infections most effectively as it treats each area as equally influential in disease transmission. Commonly, works that do consider social contact often do not consider fairness implications [**39**]. Much research considers how to optimize the age-based vaccination approach, but often other sensitive attributes are not considered [**86, 80, 48, 29, 33, 80, 37, 14**].

Similar to our approach, Minutoli et al. [**65**] lay the theoretical foundations for using IM to reduce disease transmission via vaccination, though they do not consider fairness. They use a simplistic propagation model which is not specific to a particular disease, nor does it consider important factors in disease transmission such as the number of people in a confined space, or the duration spent there, as we do. Additionally, their implementation is for contact networks containing interactions between individuals, rather than communities. We argue that privacy regulations on mobile tracking data will limit the ability to reconstruct such specific interactions, not to mention also matching demographic information to individuals in these interactions. Using aggregate visits as in our work is therefore more realistic in terms of the data available, and allows us to extract demographic information on the community level which we use for our fairness approaches. Rather than using large contact networks, condensed mobility networks of aggregate visits also drastically reduces the network size which is beneficial for computation.

Mehrab et al. use the same mobility data as in our contribution to guide vaccine distribution to groups with lower up-take rates [**61**]. However, they use visit counts to public places to determine the best candidates, whereas we determine influence by simulating disease spread on top of the mobility network to predict infections. Further, their contribution is to identify groups with low vaccine uptake at a later stage of vaccine deployment, whereas we offer a solution for the first stages of the allocation, to ensure fairness from the beginning. To our knowledge, our work is the first approach to use *community-level* influence maximization to propose vaccination strategies which consider demographic fairness.

4.3. Preliminaries

4.3.1. Influence Maximization

IM is a network science technique used to identify the most influential nodes in a graph, with respect to their ability to propagate a certain transmissible quality, such as information or disease. IM is particularly popular in viral marketing problems on social media platforms, in which an entity wants to share advertisements with only a few individuals online, but

hopes to optimize this choice by selecting those who are more likely to share the information with the rest of the network. While some techniques rely on heuristics like node centrality or betweenness to measure a node's importance in a graph [**71, 13**], IM assumes access to a function which models the propagation of the given substance across the network. The algorithm uses this propagation function to generate its selection for the set of most influential nodes, often referred to as the seed set. In their seminal work, Kempe et al. [**44**] proposed a greedy algorithm which achieves $(1 - \frac{1}{e})$ optimality guarantees on the seed set, so long as the objective function - which is the given propagation function - is both submodular and monotonic. However, the biggest drawback of the greedy strategy is its inefficient runtime. The commonly used propagation functions, such as the Linear Threshold [**21**] and Independent Cascade [**79**] models, are stochastic in nature. The greedy strategy, therefore, relies on running several Monte Carlo simulations of these stochastic propagation functions, which is costly. Much of the proceeding work following this focus on trying to improve runtime performance; CELF exploits the submodularity of the influence function to reduce the number of influence evaluations, achieving a runtime of up to 700 times faster than the greedy algorithm [**50**]. In our experiments, we also use CELF to reduce the number of evaluations.

4.3.2. Mobility networks

We draw on the work conducted by Chang et al. to construct our mobility networks and simulate Covid-19 propagation on them [**17**]. The mobility network of a metropolitan statistical area (MSA) contains K nodes which represent CBGs, neighborhoods of a few hundred to a few thousand residents. Each CBG is denoted by c_i, with $i = 1,..,K$. The population of each CBG is known and is given by n_{c_i} for CBG c_i. The total number of residents in the network is, therefore, $N = \sum_{i=1}^{K} n_{c_i}$. Each individual in the network can belong to one of M social groups (e.g. racial groups) indexed by $j = 1,..,M$. In our work, we experiment with j representing racial groups and also groups based on their median household income, obtained from US Census data [**58**]. For racial groups, each CBG can contain any number of individuals belonging to each social group, up to its total population size. The fraction of residents in CBG c_i who belong to racial group j is known and given by α_{ij}, where $0 \leq \alpha_{ij} \leq 1$, and therefore $\sum_{j=1}^{M} \alpha_{ij} = 1 \forall i$. The total number of individuals in the network belonging to racial group j is therefore obtained by $N_j = \sum_{i=1}^{K} \alpha_{ij} n_{c_i}$. For social status groups, we split the CBGs into M groups defined by the median income of the households in that CBG. The total number of individuals belonging to a group is, therefore, $N_j = \sum_{i=1}^{K} n_{c_i}$ if c_i belongs to income group j.

4.3.3. Covid-19 propagation model

We use the Covid-19 simulation proposed by Chang et al. [**17**], which models the propagation of the disease on the network of CBGs and POIs. The constructed hourly visits from the network describe how many individuals travel from CBGs to POIs per hour. POIs here represent public places such as restaurants, gyms and religious centers, where interactions with members of other communities can happen, and the disease can spread accordingly. Individuals can also transmit the disease amongst members of their home CBG. The number of new infections per CBG is therefore a summation of two terms drawn from different distributions; one Poisson distribution for new exposures from POIs, and one Binomial distribution for new exposures from their home CBG and other places which may not be accounted for in the POIs, such as public transportation. The previous work fixed the parameters of the model by comparing it to the counts of real Covid-19 cases, which we also reproduce in our work. Our vaccination approach can technically be used to combat any disease, simply by swapping the influence function σ for a model of said disease. This would mean changing the parameters of the model when calibrating it to real case counts. Further details of the model and its calibration can be found in the Appendix.

At all time steps in the disease simulation, we have access to the number of susceptible, infected, exposed, and recovered or removed individuals residing in each CBG. We maintain a vector of size K for each of these states, denoted respectively by S, E, I, and R. Their elements are indexed by i and contain the fraction of individuals in CBG c_i belonging to that state. For example, element I_i^t contains the fraction of individuals in CBG c_i who are in the infected state at time t. In this work, we are only interested in the final rates of exposed, infected or recovered/removed (EIR) individuals in the final time step of the model, T, and readers can assume we use the final iteration at T of these vectors from here onwards. For example, the sum of all exposed-or-worse individuals in the network at time T is given by $N_{EIR} = \sum_{i=1}^{K}(E_i^T + I_i^T + R_i^T)n_{c_i}$.

Individuals in the network therefore always belong to one of M social groups, as well as one of the four SEIR states. Since we only have access to these statistics on the community level, we approximate the number of people in the network belonging to racial group j and exposed-or-worse as $N_{EIR_j} = \sum_{i=1}^{K}(E_i^T + I_i^T + R_i^T)\alpha_{ij}n_{c_i}$. For income group j, the equivalent is obtained by $N_{EIR_j} = \sum_{i=1}^{K}(E_i^T + I_i^T + R_i^T)n_{c_i}$ for CBGs belonging to income group j.

4.4. Proposed Approach

In this section, we present three methods of targeted vaccination using IM. Firstly, we present our simple method for vaccinating with IM and no additional constraints. We then

outline two methods for introducing fairness to IM, for both racial groups and income groups. Finally, we present a method for applying weights to communities corresponding to their relative risk, in order to use IM and still prioritize older communities that are more vulnerable to severe outcomes if infected.

In our approach, the treatment of the three sensitive attributes (also referred to as protected attributes) in the network - race, income level and age - are not the same. For the racial and income groups, we aim to achieve fairness according to their population size in the network. An individual should be no more at risk of infection due to their race or income than what is expected given the racial or income group's population size in the network. However, the same strategy should not be adopted for age, since infection can lead to more severe outcomes for older individuals, making them more high risk. Therefore, we strive to achieve fairness amongst race and income, but adopt a *bias* with respect to age, in order to protect individuals at high-risk.

As addressed in the Preliminaries, the greedy approach in influence maximization provides provable guarantees on the optimality of the seed set, so long as the influence function is both monotonic and submodular. Previous findings have found issues proving these properties in a temporal SIR model [**25**]. Similar to this work, we argue that the greedy approach is still effective despite the propagation model exhibiting some submodularity violations. Additionally, the greedy approach is more scalable for greater network sizes than solving the optimization problem exactly, particularly when fairness constraints are required [**28**]. Further, we argue that the greedy approach provides an element of model interpretability, which is particularly important when justifying why one neighborhood should receive vaccines over another; in our case, we can clearly demonstrate with the greedy approach that a neighborhood with higher mobility and influence is more likely to be selected for vaccination. Further, we argue that using community-level influence maximization, with aggregated data on CBGs rather than individuals, is a more realistic approach when we want to obtain demographic information for fairness purposes. Privacy concerns (rightly) limit access to fine-grained data on individual mobility and their sensitive attributes, but here we use Census data to match demographic data to CBGs.

4.4.1. Vaccinating with Influence Maximization

In this section, we outline how to select the most influential communities in the network in terms of disease spread using IM and CELF. The algorithm is provided in Algorithm 2. We begin with a budget B corresponding to the number of vaccines available for allocation to the whole network. We can simulate the disease spread from a set Z of CBGs, in order to quantify how influential those communities are. In the general case, Z can contain any

number of CBGs, but to quantify the influence of just one, only that CBG would be contained in the set Z, e.g. $Z = \{c_i\}$ for $i = 1,..,K$.

For all K CBGs in the network, we then conduct simulations of the disease spreading from that neighborhood alone. We maintain a list of lists $L = [[1, \sigma(\{c_1\})],...,[K, \sigma(\{c_K\})]]$, where for CBG i, $\sigma(\{c_i\})$ represents the N_{EIR} count as a result of simulating disease spread starting only from c_i. L, therefore, contains the list of pairs of the candidate CBGs indices along with their corresponding influence, as a count of how many people resulted in exposed-or-worse states.

We initialize the set of nodes to vaccinate, Z, as an empty set. Then, in each iteration, the CBGs with the greatest marginal gain are greedily added to Z. The marginal gain is the difference in influence between the current set of selected CBGs (*spread*), and the influence of the current selected CBGs *plus a potential candidate CBG*. Note that the gain is also normalized by the population of the CBG, n_{c_i}. We implement this normalized version of IM since we want to select CBGs that are the most influential per their population, and CBGs with a higher population use more of the vaccination budget than CBGs with a lower population. We keep track of how much budget is used so far with B', which gets updated with the population sizes of CBGs when they are added to Z.

In lines 9-15, we perform a check to test whether the highest-influence candidate after the previous iteration is still the highest-influence candidate in the current iteration. If this is true, we omit the requirement to re-calculate the influence of the other candidates. This exploits the submodularity of the influence function, since the marginal gain of adding CBG c_i to a smaller set Z can only decrease. This is the contribution made by CELF [**50**], which we use to improve run time. The algorithm then continues to add candidate CBGs to Z so long as their addition does not exceed the budget B. The final output of the model is therefore a set of CBGs to be vaccinated, which we call V. Our subsequent variations of this contribution in the next sections adapt this method to apply demographic fairness.

4.4.2. IM with equal treatment

Equal treatment is an existing fairness notion in the domain of fair IM, which aims to achieve fair representations of social groups in the final set of selected nodes V. This is equivalent to achieving the same demographic distribution in the set of communities to be selected for vaccination as in the whole network. We model this task as a multiple knapsack problem [**18**], whereby each social group $j = 1,..,M$ is allocated a number of vaccines based on the fraction of their population in the network. Each social group j, therefore, receives its own budget B_j, corresponding to the number of vaccines to be allocated to the group,

Algorithm 2 Selecting CBGs to vaccinate using IM and CELF

1: **Input**: budget B, number of CBGs K, disease model σ
2: **Output**: CBGs to vaccinate V
3: $B' \leftarrow 0, spread \leftarrow 0, L \leftarrow [], Z \leftarrow []$
4: **for** $i = 1, \ldots, K$ **do**
5: $\quad L$.append$([i, (\sigma(\{c_i\}) - spread)/n_{c_i}])$
6: **end for**
7: sort L by gain, descending
8: Z.append$(L_{0,0})$ {add cbg with best gain to Z}
9: $spread \leftarrow L_{0,1} * n_{Z_{-1}}$
10: $B' \leftarrow n_{Z_{-1}}$ {update budget used}
11: **while** there are possible candidates in L **do**
12: $\quad matched \leftarrow False$
13: $\quad$ **while** not matched **do**
14: $\quad\quad best \leftarrow L_{0,0}$
15: $\quad\quad L_{0,1} \leftarrow (\sigma(Z \cup \{best\}) - spread)/n_{best}$
16: $\quad\quad$ sort L by gain, descending
17: $\quad\quad matched \leftarrow L_{0,0}$ is $best$
18: $\quad$ **end while**
19: $\quad spread \leftarrow spread + L_{0,1} * n_{best}$
20: $\quad Z$.append$(L_{0,0})$
21: $\quad B' \leftarrow B' + n_{Z_{-1}}$ {update budget used}
22: $\quad L \leftarrow L[1:]$ {remove best from the candidate list}
23: $\quad$ keep only candidates in L which cannot exceed B
24: **end while**
25: **return** Z

given by

$$B_j = \frac{N_j}{N} B \tag{4.4.1}$$

where N_j is the number of individuals in the network belonging to group j and N is the total network population. To implement this, when a CBG is selected for vaccination, we update the budget used by each of the M social groups, between lines 19 and 20 of Algorithm 2. Additionally, after line 21, we perform another check to ensure the remaining candidate list contains only CBGs whose addition would not violate any of the social group budgets B_j. We use these definitions to outline two strategies: equal treatment by racial groups and equal treatment by median household income.

4.4.2.1. Equal treatment by racial groups (IM-R). When performing equal treatment for racial groups, the budget of racial group j is updated when CBG c_i is selected for vaccination via

$$B'_j = B'_j + n_{c_i} \alpha_{ij} \tag{4.4.2}$$

This update is performed for all M racial groups when any additional CBG is selected.

We acknowledge that our fairness framework takes a Western-centric perspective, particularly with respect to race, and is relevant mostly to countries containing urban areas with high diversity, as is more typical of the global West. The most effective strategy for each country, however, will not be the same [**26**]. We therefore propose IM-I to counteract this bias; many non-Western countries may not have the same extent of racial diversity, but will still experience income disparities in their urban areas. Using income as a sensitive attribute therefore still provides a fair IM method which is relevant for areas with low racial diversity. Further, Figure 4.2 identifies that in our selected MSAs, the White population tends to dominate higher income CBGs, while historically marginalized groups of Black or African-American are more prevalent in lower income groups, as is typical of high-diversity urban areas in the West. Therefore, fairness by income level may also achieve fairness by racial groups for our selected MSAs.

4.4.2.2. Equal treatment by median income (IM-I). We perform a similar equal treatment scenario, this time with social groups defined by income. We use labels of the median household income of each CBG. The distribution of the CBG median income is split into four quartiles. We then bucket the CBGs into one of $M = 4$ groups according to which quartile its median income falls into. The budget is split in the same way as with race, using $B_j = \frac{N_j}{N} B$ for each income group j, where N_j is the total population of that income group. However this time, the budget updates are given by

$$Bj' = \begin{cases} B'_j + n_{c_i}, & \text{if } c_i \text{ belongs to group } j \\ B'_j, & \text{otherwise} \end{cases} \tag{4.4.3}$$

4.4.3. IM with age-associated risk-weights (IM-A)

While older individuals in the network are less mobile, they are more likely to experience severe consequences if they contract the disease, including hospitalisation and death. As a result, it is important to consider this tradeoff when using IM. To do so, we incorporate this notion into our IM technique such that the significance of infecting a person from a CBG with a higher median age is greater. We implement this by weighting the CBGs with a "risk-factor" according to their median age. The purpose is to not only select CBGs for vaccination that are highly influential in the network but in particular, find those CBGs that pose more of a risk of exposing older communities to the disease. We achieve this by scaling the influence calculations σ, which is used as the selection criteria for vaccination (see Algorithm 2). We construct a vector μ of size K containing the associated risk-weights for each CBG, determined by its median age. We fill this vector with the death rates per age group from the CDC, which are the rates of death of older groups compared to 18-29

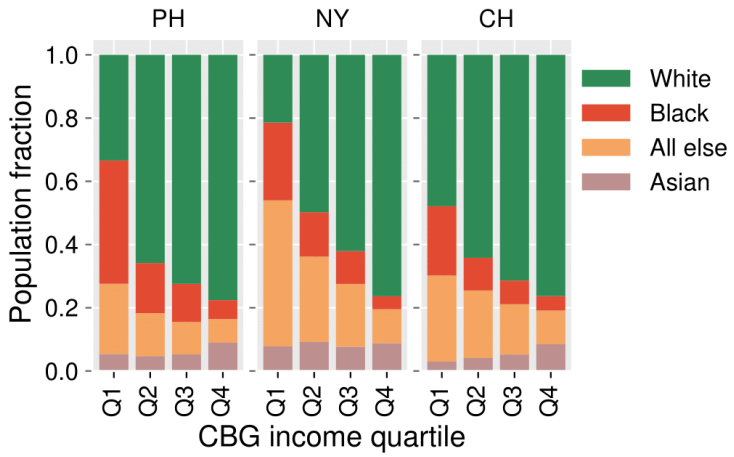

Fig. 4.2. Racial distributions of CBGs grouped by their median income. Income groups are determined by quartiles of the median income distribution. Results are for three MSAs: Philadelphia, New York and Chicago.

year olds in the US [**16**]. We report these values in the Appendix. The resulting influence, which before was just the sum of exposed-or-worse individuals in the network, is now instead a weighted sum where each CBG is weighted according to its risk-factor in vector μ, given by

$$\sigma_A = \sum_{i=1}^{K} \mu_i n_{c_i} (E_i + I_i + R_i). \tag{4.4.4}$$

We use this metric as a proxy to account for infections as well as more severe cases and deaths. In Algorithm 2, the influence function σ is replaced by σ_A in lines 3 and 13.

4.4.4. Multiple Protected attributes

The methods proposed so far each strives for fairness using one protected attribute at a time. However, it is possible that achieving the desired fairness for one protected attribute is not beneficial for another. For example, older communities tend to be predominantly White in the US. Therefore, a vaccination strategy based solely on age, without consideration for fair distribution amongst racial groups, leads to an unfair allocation. This would also lead to unfair allocation amongst income groups as White populations tend to dominate higher income CBGs, as shown in Figure 4.2. To address this issue, we propose the following combinations of our methods:

4.4.4.1. IM- with Race groups and Age-associated risk-weights (IM-RA). We perform equal treatment to achieve representative allocation of vaccines amongst racial groups, but use the influence function σ_A to apply a heavy penalty for infecting older communities.

4.4.4.2. IM - with median Income and Age-associated risk-weights (IM-IA). Similarly, we perform equal treatment of vaccines amongst the four income groups and replace the influence function with σ_A. In our experiments, we later find high numbers of submodularity violations when using the σ_A influence function. For all experiments using the "-A" suffix, we, therefore, omit the recalculation checks in Lines 10 to 15 of Algorithm 2.

4.5. Experiments

Dataset We conduct experiments on three mobility networks of MSAs in the US, constructed from individual mobile tracking data from SafeGraph, from the Dewey platform [**24**]. We use the implementation proposed by [**17**] to construct these networks, and also use their Covid-19 model as our influence function. A mobility network is constructed as a temporal bipartite graph $G^t = (V,E^t)$, whose nodes V are either CBGs, which are communities of between 600 and 3,000 US residents, or POIs. A directed weighted edge $e^t(v_1,v_2) \in E^t$ represents the number of residents from a CBG who are visiting a POI at hour t. The graph varies over time, such that no nodes are added or removed, but the edge weights vary hourly. We use five weeks of visits from CBGs to POIs beginning on the 2nd of March 2020. For our implementation, only the first two weeks are used to select the most influential communities, and vaccination is implemented after the two-week mark. We used data from the beginning of March in order to exploit the full mobility of individuals before lockdown. Otherwise, it would not have been possible to measure the effect of the proposed vaccination strategies separately from the effect of lockdown. We include details of the computing infrastructure for creating the networks and running experiments in the Appendix.

Throughout the experiments, we set the vaccination budget to 5% of the population size of the whole network. We experiment with three MSAs - Philadelphia, New York and Chicago - each of which encompasses the main city as well as the wider metropolitan area. These particular MSAs were selected based on their high racial diversity, and high discrepancy in infections between racial groups as reported in the previous work. We provide more details of this setup, including further pre-processing steps, in the Appendix. The outputs are mobility networks modelling metropolitan areas of populations between 6 and 10 million residents. We provide statistics of network size after these pre-processing steps in the Appendix.

Baselines Each vaccination strategy is run over 30 random seeds, and we report the average results in comparison to three baseline approaches. We design our own baselines here, since the closest approach to ours which uses IM for vaccination uses a generic propagation function

which is not specific to a particular disease, and is valid only on rooted trees and not general graphs [**65**].

- *No vaccination:* We model free-spreading Covid-19 during the total five-week period, without any vaccination strategy.
- *Random vaccination (RAND):* We implement a random selection of CBGs for vaccination within the budget B. We collect results over three random seeds and report the average.
- *Current strategy proxy (CS):* Here we replicate the current strategy which prioritizes older communities. We select the oldest communities for vaccination, by the median age of the CBG, up to the vaccine budget B.

Fairness Evaluation Metrics In addition to evaluating the performance of the vaccination strategies, we propose two methods of evaluating fairness for social groups in this context. In both cases, we measure the discrete KL-divergence between two distributions; we compare a distribution from the outcome of our experiments $p(j)$ to an ideal "fair" distribution $q(j)$. We draw on fairness notions from the IM literature - equal treatment and equal outcomes. For both measures, the fair distribution $q(j)$ corresponds to the fractions of each social group in the network, $q(j) = N_j/N$.

- **Equal treatment** For equal treatment, we aim to obtain a fair representation of each social group j within the CBGs selected for vaccination. As such, the output distribution $p(j)$ is the proportion of social group j amongst vaccinated CBGs, $p(j) = (N_j/N)_V$.
- **Equal outcome** To obtain equal outcomes, the goal is to ensure that no individual is more at risk of infection than the fraction of that social group in the network dictates. In this case, we set $p(j)$ to be the proportion of infections received by social group j, as a result of our vaccination strategy. This can be written as $p(j) = (N_j/N)_{EIR}$.

4.6. Results and Discussion

4.6.0.1. Reducing overall infections. The performance of each vaccination strategy in reducing the number of infections is shown in Figure 4.3 (green). For every MSA, all vaccination strategies using IM outperform both RAND and CS. Though the infections decrease by only a few percent, given the network size, these percentage point differences are significant. For example, a 5% decrease in infections for Philadelphia corresponds to around 18,000 fewer people infected (including estimates of unreported infections). We observe that the variations of IM experiments which include fairness (the last five bars) do not experience a significant decrease in performance even when optimizing for both performance *and* fairness.

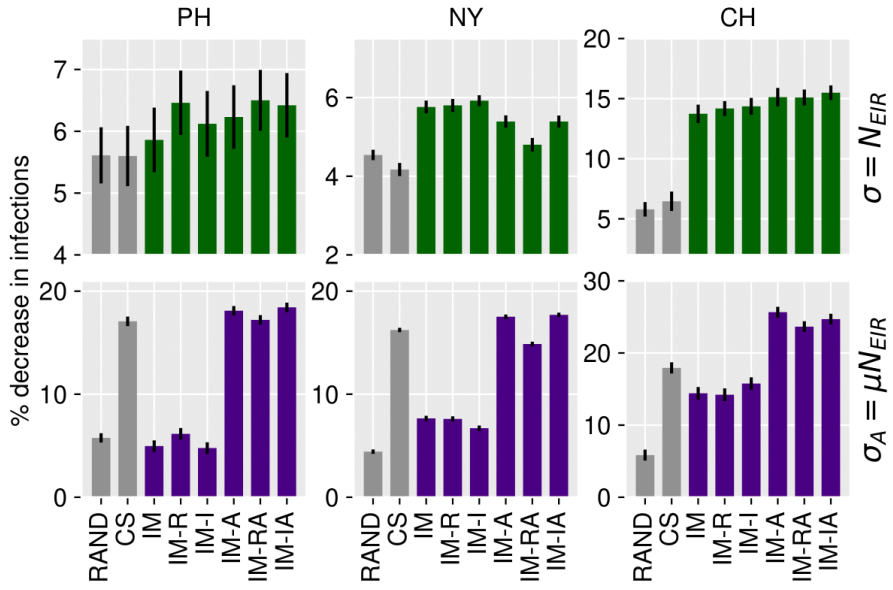

Fig. 4.3. Performance measured by percentage decrease in infections (top), and percentage decrease in risk-weighted infections, i.e. with a weighted penalty of infecting older communities (bottom), compared to not vaccinating. Higher is better for both metrics.

This illustrates that fairness considerations in vaccination distribution do not have to come at the cost of increasing infection counts.

4.6.0.2. Infections in high-risk groups. Figure 4.3 (purple) reports the percentage decrease in infections weighted by age-associated risk, as described in Equation 4.4.4. The experiments optimizing for age are the baseline CS and our contributions IM-A, IM-RA, and IM-IA, which therefore perform best for this metric. However, we see that all of our proposed solutions which optimize for age outperform the current strategy, even when they are also optimizing for another sensitive attribute at the same time. The results testify to several alternative solutions that better protect older communities as well as ensure fairness for other sensitive attributes like race and income.

4.6.0.3. Comparing fairness notions. In Figure 4.4 left, we present results for equal treatment (red) and equal outcome (blue) for racial groups, and in Figure 4.4 right we present the same metrics for income groups. Though the D_{KL} values are small, these still correspond to significant differences in these large networks. For example, the current strategy model (CS) has a D_{KL} of around 5.8×10^{-4} for Philadelphia, which corresponds to the Black population suffering around 6,000 more infections than if they were distributed according to their proportion in the network. Critically, we see that IM-R methods successfully achieve equal treatment for racial groups across all MSAs, as does IM-RA, with the exception of

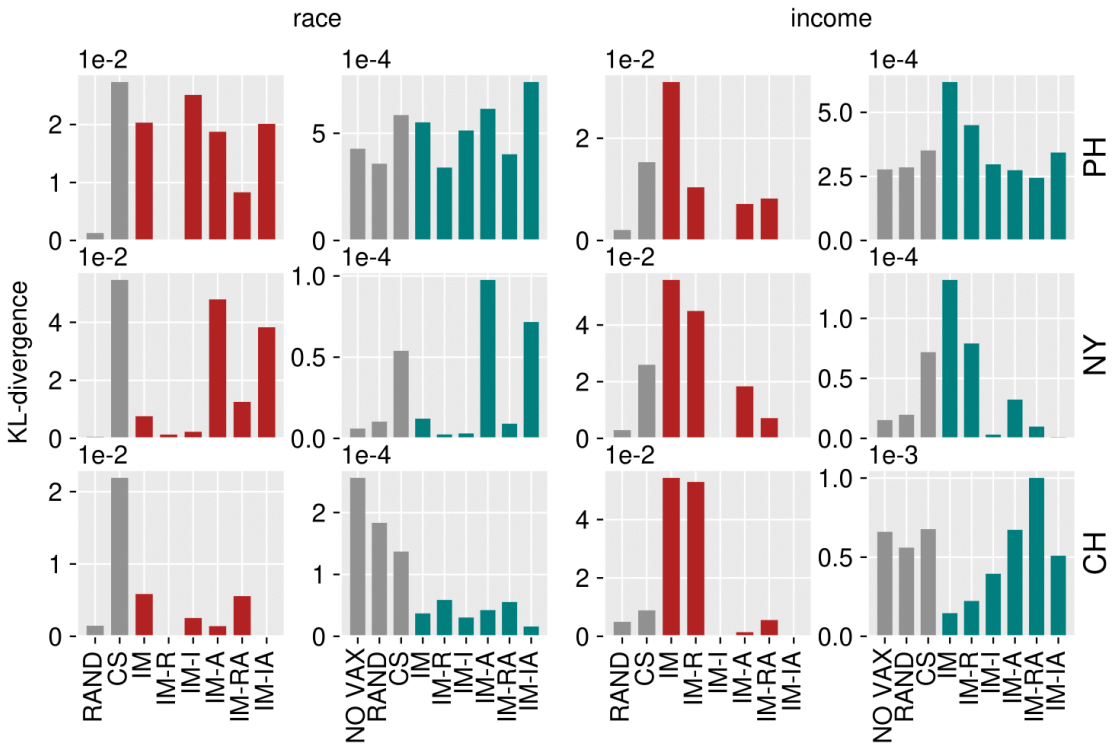

Fig. 4.4. The KL-divergence scores measure fair treatment (red) and fair outcomes (blue) with respect to racial groups (left) and income groups (right). Lower D_{KL} corresponds to better fairness for both metrics.

Chicago. The effect is stronger for income, where IM-I and IM-IA experiments achieve near-perfect (D_{KL} of zero) distribution of vaccines to income groups, for all MSAs. Though equal treatment is important to ensure, we are more interested in an equal outcome, as infections are the more serious consequence. We see that the experiments which perform best for equal outcomes amongst races are also IM-R and IM-RA. This demonstrates that achieving equal treatment can be very effective in delivering fair outcomes for those same demographic groups. Additionally, for some MSAs, constraining on fairness by *income* (IM-I) leads to fair outcomes for *race*, and vice versa. This implies that the objectives of achieving fairness amongst races and fairness amongst social status are similar.

4.6.0.4. Optimizing for multiple sensitive attributes. Since the results of every vaccination experiment for each MSA can differ, it is possible that there is no one one-size-fits-all best vaccination strategy for every urban area. In particular, we can identify how accounting for higher-risk individuals (experiments with the "-A" suffix) can work favourably for achieving demographic fairness in infections for some but not all MSAs (see Chicago, Figure 4.4 blue, left and New York, Figure 4.4 blue, right). Despite this, we can identify at least one strategy per metropolitan area which achieves high performance in reducing infections

overall, as well as a competitive result for the infection outcomes of *all three* sensitive attributes of age, race, and social status by income: IM-RA for Philadelphia and New York, and IM-IA for Chicago.

4.7. Conclusion

Fair vaccination strategies are essential to protect at-risk communities and mitigate disparities among different social groups. As minorities and lower-income communities were not able to reduce their movement during the pandemic, we proposed alternative strategies based on the mobility of individuals using influence maximization.

For policy-makers, choosing a vaccination strategy amongst those presented here is non-trivial. There is no one-size-fits-all solution for every urban area. However, we demonstrate that, for all networks we tested here, one of our proposed methods can successfully ensure demographic fairness for all three sensitive attributes. We, therefore, argue that community-level influence maximization should be incorporated into whichever ethical stance is taken, and we present the methodology to do so.

Our approach can be extended to accommodate multiple rounds of vaccine allocation, as commonly observed in real-world scenarios. It would be necessary to capture the mobility shifts at different stages of lockdown, and how this affects demographic groups differently. There are many factors which could be incorporated to make the simulation more realistic, such as vaccine hesitancy. For later stages of vaccine roll-out, our approach could be combined with data on rates of uptake or hesitancy to influence the selection of neighborhoods.

4.8. Appendix

4.8.1. Constructing Mobility Networks

We use the Covid-19 simulation and mobility network proposed by Chang et al. [**17**]. To construct the networks, they used the Safegraph social-distancing data of daily estimates of the fraction of CBG's residents who are out visiting other CBGs. Safegraph no longer provides these values, so we estimated it ourselves using Safegraph's Neighborhood Patterns of visitors *arriving at* CBGs, and aggregating over these. We make our code available upon publication of this work to demonstrate our approach. Additionally, due to sparsity in the weekly patterns data used to construct the mobility networks, we aggregate over two previous months of the monthly patterns data. We use fewer months of aggregate visits than the previous approach, which means our networks' statistics differ from theirs. This also leads to Covid-19 model parameters which differ from that of the previous work (see Table 4.1).

	β_{home}	ψ	p_0
Philadelphia	0.02	300	0.001
New York	0.02	100	0.005
Chicago	0.02	500	0.0005

Table 4.1. Final model parameters obtained by tuning infection model to real Covid-19 case counts for each MSA.

Table 4.2 summarizes the network statistics of the three MSAs in terms of population, CBGs, and POIs.

MSA	Population	CBGs	POIs
Philadelphia	9,247,281	5,603	11,479
New York	9,990,617	6,522	20,606
Chicago	6,074,364	3,452	8,281

Table 4.2. Final statistics of mobility networks.

4.8.2. Covid-19 Model

Below we describe the basic set-up of the Covid-19 model. More details can be found in the supplementary information of the previous work by Chang et al. [**17**]
The model maintains four vectors of size K - S, E, I and R - corresponding to the fraction of susceptible, exposed, infected and recovered/removed individuals per CBG. The rates of transitions between these states at a time step t are determined by the number of new

exposures $N^{(t)}_{S_{c_i} \to E_{c_i}}$, the number of exposures transitioning to infections $N^{(t)}_{E_{c_i} \to I_{c_i}}$, and the number of infections transitioning to recovered/removed, $N^{(t)}_{I_{c_i} \to R_{c_i}}$. The number of new exposures depends on two factors; visits to public places (from the visit matrix w) containing other infectious individuals, and interactions from the home CBG with other infectious individuals. Transitions are sampled from the following distributions:

$$N^{(t)}_{S_{c_i} \to E_{c_i}} \sim \text{Pois} \frac{S^{(t)}_{c_i}}{N_{c_i}} \sum_{j=1}^{n} \lambda^{(t)}_{p_j} w^{(t)}_{ij} + \text{Binom}(S^{(t)}_{c_i}, \lambda^{(t)}_{c_i}) \tag{4.8.1}$$

$$N^{(t)}_{E_{c_i} \to I_{c_i}} \sim \text{Binom}(E^{(t)}_{c_i}, 1/\delta_E) \tag{4.8.2}$$

$$N^{(t)}_{I_{c_i} \to R_{c_i}} \sim \text{Binom}(I^{(t)}_{c_i}, 1/\delta_I) \tag{4.8.3}$$

Here, λ_{p_j} refers to the infection rate at POI p_j, and λ_{c_i} is the infection rate at CBG c_i. δ_E and δ_I refer to the mean exposure period and the mean infectious period respectively.

Using the number of transitions, the number of S, E, I and R for a CBG c_i at time $t > 0$ can be expressed as follows:

$$S^{(t)}_{c_i} = S^{(t-1)}_{c_i} - N^{(t)}_{S_{c_i} \to E_{c_i}} \tag{4.8.4}$$

$$E^{(t)}_{c_i} = E^{(t-1)}_{c_i} - N^{(t)}_{E_{c_i} \to I_{c_i}} + N^{(t)}_{S_{c_i} \to E_{c_i}} \tag{4.8.5}$$

$$I^{(t)}_{c_i} = I^{(t-1)}_{c_i} + N^{(t)}_{E_{c_i} \to I_{c_i}} - N^{(t)}_{I_{c_i} \to R_{c_i}} \tag{4.8.6}$$

$$R^{(t)}_{c_i} = R^{(t-1)}_{c_i} + N^{(t)}_{I_{c_i} \to R_{c_i}} \tag{4.8.7}$$

At $t = 0$:

$$E^{(0)}_{c_i} = \begin{cases} N_{c_i} p_0 & \text{if } c_i \text{ is infected} \\ 0 & \text{otherwise.} \end{cases} \tag{4.8.8}$$

$$S^{(0)}_{c_i} = N_{c_i} - E^{(0)}_{c_i} \tag{4.8.9}$$

$$I^{(0)}_{c_i} = 0 \tag{4.8.10}$$

$$R^{(0)}_{c_i} = 0 \tag{4.8.11}$$

where p_0 is the probability that an individual in the metro area is exposed at the first time step. This probability is one of the three parameters set when calibrating the disease propagation to real case counts, see Table 4.1.

4.8.3. Age group risk factors

Table 4.3 shows the different risk factors of infection, hospitalisation, and death of different age brackets compared to 18-29 years olds. The trend shows that while younger people are more prone to getting infected, they are less likely to face grave or fatal consequences.

For instance, a person aged 85 is 350 times more likely to die from covid compared to a person in their 20s. It is important to consider such risks when using developing a vaccination strategy.

	30-39	40-49	50-64	65-74	75-84	85+
Cases	1.0x	0.9x	0.8x	0.6x	0.6x	0.7x
Hosp.	1.5x	1.9x	3.1x	4.8x	8.6x	15x
Death	3.5x	10x	25x	60x	140x	350x

Table 4.3. Age group risk factors of cases, hospitalization and death compared to 18-29 year olds, from CDC [**16**].

4.8.4. Mobility of selected CBGs for vaccination

Table 4.4 contains the average mobility of CBGs who were selected for vaccination, both before lockdown and during lockdown. The first column contains the average pre- and in-lockdown mobility of the entire network. Strategies which select higher mobility CBGs to vaccinate generally have better performance in reducing overall infections (see Figure 4.3, green).

		NTWRK	RAND	CS	IM	IM-R	IM-I	IM-A	IM-RA	IM-IA
PH	pre-lockdown	2.069	2.066	1.956	1.986	**2.149**	2.009	2.042	2.077	2.061
	in-lockdown	0.830	0.822	0.779	0.822	**0.897**	0.838	0.806	0.832	0.812
NY	pre-lockdown	2.956	2.997	2.835	**3.111**	**3.111**	3.087	2.966	2.968	2.909
	in-lockdown	0.973	0.966	0.992	**1.049**	1.039	1.014	1.007	0.994	0.990
CH	pre-lockdown	2.108	2.097	2.034	2.213	**2.246**	2.151	2.119	2.137	2.141
	in-lockdown	0.763	0.764	0.745	0.781	**0.819**	0.796	0.758	0.777	0.795

Table 4.4. The average mobility of CBGs selected for vaccination for each strategy, compared to average mobility in the whole network (NTWRK). Values for pre-lockdown and in-lockdown are presented, with the highest mobility values per row in bold.

4.8.5. Infrastructure

To first construct the three mobility networks, we use 2CPUs and 64Gb of memory on Linux OS. The vaccination experiments can then each be run on 1CPU, 2GPU and 8Gb memory. We point the reader to the README file included in the code repository for a list of libraries and packages required to run on conda environments. The code can be found here: (we make this link live upon publication).

Chapter 5

Conclusion and Future Work

In this work, we present our contributions for tackling the problem of unequal protection from disease outbreaks for economically-disadvantaged groups and racial minorities using a vaccination deployment strategy which considers demographic fairness. Using aggregate statistics of visits to public places, we first optimize for the overall reduction in infections using an influence maximization approach. We then build on this basis to apply constraints equivalent to fair access to vaccines for racial groups and income groups, and a prioritization for older groups due to their increased risk of severe symptoms. Our proposed techniques are combinations of these demographic fairness considerations, whereby we optimize for several sensitive attributes at once.

Amongst the solutions we present in this work, it is evident that due to their differing mobility patterns and demographic configurations, each MSA has a different best approach. If we take the prioritization of older groups as an absolute requirement, respecting the obligations which governments currently abide by, we are left with three of our IM solutions to choose from per MSA; IM-A, influence maximization with age priority, IM-RA, age priority and racial group fairness, and IM-IA, age priority and income group fairness. Of these three, the best performing technique is different for each MSA when we take into consideration all the performance and fairness metrics by which we evaluate them. When selecting from these three, we find that their order is similar for fair treatment and fair outcomes per MSA; for example in Figure 4.4, in New York, based on fair treatment for races (red, left) the clear best choice is IM-RA, followed by IM-IA and then IM-A. We observe the same pattern in the fair *outcomes* for race in New York (blue, left). This pattern is replicated in all of the treatment-outcome plot pairs (except for income in Philadelphia). This is further evidence that, although we only implement fairness in terms of equal treatment for our sensitive attributes (race and income), this leads to the desired equal outcome. Put simply, fair

distribution of vaccines with respect to a sensitive attribute leads to fair rates of infection for that sensitive attribute *without needing to explicitly constrain on it.*

It is also important to note that fair treatment (red) and fair outcome (blue) measures do not always coincide (see Figure 4.4); in particular, for income fairness in Chicago, the plain IM approach is most unfair in terms of treatment but most fair in terms of outcome. Indeed, the shapes of the red and blue income plots for Chicago are strikingly dissimilar, whereas for New York and Philadelphia (above) we could say there is an approximate similar trend between treatment and outcome. This demonstrates the importance of evaluating each MSA separately, as it is the consequence of their mobility dynamics and their unique patterns in wealth and race distribution in the network which affects how each solution will perform. In practise, it would be the responsibility of local governments to implement the framework and oversee this choice.

5.1. Fairness-performance trade-off

Based on the typical behaviour of fair models studied in fairness literature, we would expect the plain IM approach to perform best in terms of reducing overall infections, since additional fairness considerations may introduce a performance-fairness trade-off. However, our results demonstrate that some of our approaches which include fairness perform better in terms of reducing overall infections than the plain IM approach. There is an explanation for this phenomenon based on our choice of time period in which we test and evaluate our models. The period in which we select CBGs occurs before lockdown was introduced, however our subsequent evaluation period does overlap with lockdown. Our initial observations revealed that minority communities exhibited less capacity to reduce their mobility during lockdown, as demonstrated in Figure 4.1. With this discovery, we anticipation that minority communities might be more likely to be selected for vaccination if our selection period coincided with the lockdown. The objective was to demonstrate that even if the plain IM strategy inherently disadvantaged minority communities, our fairness-enhancing method could rectify this bias. By conducting this test before the lockdown, we established a compelling case for our model's fairness, thereby showcasing its effectiveness in addressing disparities in infections.

So while the plain IM approach does not actively favour minorities, picking the most mobile CBGs while everyone's mobility was at its highest, it is evaluated during a time when minorities are more mobile. It is therefore possible that the fairness-enhanced method performs better during the evaluation since minorities were more mobile. In practice, we may not witness this same effect since both the selection and evaluation periods would both occur

in lockdown. As such, the model should be applied to mobility data immediately before vaccine availability. This would likely result in a more diverse selection of CBGs for vaccination, as minorities would exhibit less mobility reduction. Nevertheless, we intentionally chose to test the model during a period when the trade-off between performance and fairness is more pronounced, posing a greater challenge to the model's capabilities.

5.1.1. Defining fairness

It is an interesting line of argument to consider the following question: what about how the infection rates disproportionately affected economically-disadvantaged groups and racial minorities *before* the vaccine was available? Is it the responsibility of the vaccine deployment orchestrators to make up for this, by providing a greater advantage to those already disadvantaged from the period spanning the beginning of the outbreak until the availability of the vaccine? One could argue this case by citing the importance of equity. In this work, we instead only consider fairness from the moment the vaccine becomes available, as a "clean-slate". If policy-makers wanted to implement a positive bias towards the disadvantaged group(s) who suffered more infections and deaths than their population size would account for, then this would be possible to implement within our framework, simply by changing the "fair" distributions in order to up-weight the previously marginalized groups. However, the new distribution choice would have to be rigorously justified, in order to place a numerical value on how much positive bias and extra protection these groups should inherit. This is, again, a very non-trivial ethical problem, but made possible in our framework should it be considered by policy-makers.

5.2. Limitations

One of the biggest limitations of the solution presented here is the use of aggregate-level age data. Having only statistics of the median age of CBGs means that we can only penalize a method for infecting an individual from a CBG with a high median age. This means there are many older people who will not be accounted for, living in CBGs with lower median ages, but the model is not penalized for infecting them. A potential work-around could be found by utilizing the hesitant populations of each CBG. In each CBG, there will be some members who display vaccine hesitancy, and while our model will allocate vaccines to all members of the CBG if the CBG is selected, some of these will go unused if the uptake rate is not 100%. While it is hard to predict hesitancy rates for the first round of vaccination, we can imagine a re-allocation scheme where we anticipate a certain hesitancy rate in all CBGs, and those vaccines, which would otherwise go unclaimed, are re-distributed to older members of other CBGs. In this sense, it is possible to merge our

solutions even more so with the current strategy of vaccinating older individuals first. Although this may introduce a degree of realism in estimating the impact of hesitancy on total infections, a uniform hesitancy rate for all CBGs would not affect the selection process for CBGs, particularly given that vaccination is only introduced during the evaluation period.

Using a SEIR model as in our case, there are many ways to adapt the model to be ever more realistic. Firstly, in this work we model vaccination by setting the whole CBG to reach full immunity, reducing their rate of transmission to zero. We also implement this full immunity immediately after the vaccine is available. As we know, Covid-19 vaccines were never 100% effective, required several weeks to reach their highest immunity rate, and their efficacy reduced for older recipients [**27, 54**]. In this sense, a more realistic model could implement an efficacy rate drawn from some random distribution, based on the average age of the CBG, as well as increasing the efficacy gradually over the first few weeks. This also brings us to the limitations of the demographic information: the age data and income level data are simply averages over the CBG neighborhood, and the race data is also the racial distribution on the aggregate CBG level. In this sense, when we report final infection rates for racial groups, we are limited to sampling that racial group from the final infections rates of the CBG. The same can be said for calculating mobility. Despite these limitations, we were still able to extract the expected trends from the data which matched what we know about how wealth is distributed among racial groups in the US (Figure 4.2) and in who was able to reduce their mobility quicker (Figure 4.1). This reassures us that our results with respect to infections in demographic groups are also extracting reliable trends despite only having access to aggregate-level data.

5.3. Future work

The vaccine distribution solutions proposed in this work are tested on only five weeks of mobility data. As such, there is potential to extend the proposed methods to facilitate several rounds of vaccine roll-out. In reality, the first round of vaccines were not all distributed at once. It is therefore an on-going problem of optimizing the distribution process when new rounds of the vaccine become available. In order to extend our solutions to fit this problem, it would be necessary to map simulations to on-going infection rates in order to calibrate the model to real cases more than just once. Considerations of who has already received the vaccine, and how high the up-take rates are per CBG are important factors and would add extra layers of complexity. There exists work focusing on targeting communities of high hesitancy and low up-take, after initial rounds of deployment have already occurred; pairing

this with our optimization and fairness methods using influence maximization would make for an interesting direction for a sequential problem.

We also explored a method of constraining explicitly on the infections in order to directly achieve fair outcomes in terms of how infections were distributed amongst demographic groups. Inspired by previous work in fair IM, we experimented with the use of a regularizer to explicitly constrain on the infections [**7**]. This was an alternative to our equal treatment approach proposed in the paper, which only deals with how the vaccines themselves are distributed rather than the infections in the outcome. Though we did not have success with our attempts of using a regularizer, we believe there is still potential for this direction.

While the specific propagation model we use exhibited some submodularity violations, it would be a significant contribution to design a completely submodular propagation model to simulate infections on these same types of mobility networks. While the greedy approach we use here is effective compared to baseline methods - despite our influence function showing some submodularity violations - one could provide explicit optimality guarantees with use of an entirely submodular function. Such guarantees are common in IM literature, where approaches are often evaluated based on their proximity to the optimal solution. We made attempts to prove submodularity of our influence function theoretically, but the proof became intractable when iterating over hundreds of hours of simulation equations, particularly as the graph edges vary hourly. Either the development of an alternative submodular spread function, or identifying the optimality guaruntees with the function used in this work, would be valuable contributions to the research direction and strengthen the robustness of this approach.

References

[1] Ritu Agarwal, Michelle Dugas, Jui Ramaprasad, Junjie Luo, Gujie Li, and Guodong (Gordon) Gao. Socioeconomic privilege and political ideology are associated with racial disparity in COVID-19 vaccination. *PNAS, Proceedings of the Nation Academy of Sciences of the United States of America*, 118, 2021.

[2] Donald J. Alcendor. Racial Disparities-Associated COVID-19 Mortality among Minority Populations in the US. *Journal of Clinical Medicine*, 9, 2020.

[3] Donald J. Alcendor, Paul D. Juarez, Patricia Matthews-Juarez, Sheena Simon, Catherine Nash, Kirollos Lewis, and Duane Smoot. Meharry Medical College Mobile Vaccination Program: Implications for Increasing COVID-19 Vaccine Uptake among Minority Communities in Middle Tennessee. *Vaccines*, 10, 2022.

[4] Junaid Ali, Mahmoudreza Babaei, Abhijnan Chakraborty, Baharan Mirzasoleiman, Krishna Gummadi, and Adish Singla. On the Fairness of Time-Critical Influence Maximization in Social Networks. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[5] Hadis Anahideh, Lulu Kang, and Nazanin Nezami. Fair and Diverse Allocation of Scarce rResources. *Socio-Economic Planning Sciences*, 80, 2022.

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. *ProPublica*, 2016.

[7] Md Sanzeed Anwar, Martin Saveski, and Deb Roy. Balanced Influence Maximization in the Presence of Homophily. *WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 175–183, Aug 2021.

[8] Kristen M.J. Azar, Zijun Shen, Robert J. Romanelli, Stephen H. Lockhart, Kelly Smits, Sarah Robinson, Stephanie Brown, and Alice R. Pressman. Disparities in Outcomes Among COVID-19 Patients in a Large Health Care System in California. *Health Affairs*, 39(7):1253–1262, jul 2020.

[9] Mohsen Bayati, Rayehe Noroozi, Mohadeseh Ghanbari-Jahromi, and Faride Sadat Jalali. Inequality in the distribution of Covid-19 vaccine: a systematic review. *International Journal for Equity in Health*, Aug 2022.

[10] Dimitris Bertsimas, Vassilis Digalakis Jr, Alexandre Jacquillat, Michael Lingzhi Li, and Alessando Previero. Where to locate COVID-19 mass vaccination facilities? *Naval Research Logistics*, 69, 2021.

[11] Dimitris Bertsimas, Joshua Ivanhoe, Alexandre Jacquillant, Michael Li, Alessandro Previero, Omar Skali Lami, and Hamza Tazi Bouardi. Optimizing Vaccine Allocation to Combat the COVID-19 Pandemic. *medRxiv [preprint]*, 2020.

[12] Philip Blumenshine, Arthur Reingold, Susan Egerter, Robin Mockenhaupt, Paula Braveman, and James Marks. Pandemic Influenza Planning in the United States from a Health Disparities Perspective. *Emerging Infectious Diseases*, 2008.

[13] Stephen P. Borgatti. Centrality and Network Flow. *Social Networks*, 27:55–71, 2005.

[14] Jack H. Buckner, Gerardo Chowell, and Michael R. Springborn. Dynamic prioritization of COVID-19 vaccines when social distancing is limited for essential workers. *Proceedings of the National Academy of Sciences*, 2021.

[15] Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR*, 2018.

[16] CDC. Risk for COVID-19 Infection, Hospitalization, and Death By Age Group. `https://archive.cdc.gov/#/details?url=https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-age.html`, 2023. Last Accessed: 2023-01-24.

[17] Serina Chang, Emma Pierson, Pang Wei Koh, Jaline Gerardin, Beth Redbird, David Grusky, and Jure Leskovec. Mobility network models of COVID-19 explain inequities and inform reopening. *Nature 2020 589:7840*, 589(7840):82–87, Nov 2020.

[18] Chandra Chekuri and Sanjeev Khanna. A Polynomial Time Approximation Scheme for the Multiple Knapsack Problem. *SIAM Journal on Computing*, 35, 2005.

[19] Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010.

[20] Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. *KDD*, 2009.

[21] Wei Chen, Yifei Yuan, and Li Zhang. Scalable Influence Maximization in Social Networks under the Linear Threshold Model. *IEEE International Conference on Data Mining*, 2010.

[22] Suqi Cheng, Huawei Shen, Junming Huang, Guoqing Zhang, and Xueqi Cheng. StaticGreedy: Solving the scalability-accuracy dilemma in influence maximization. In *International Conference on Information and Knowledge Management, Proceedings*, 2013.

[23] Erica Jane Cook, Elizabeth Elliot, Alfredo Gaitan, Ifunanya Nduka, Sally Cartwright, Chimeme Egbutah, Gurch Randhawa, Muhammad Waqar, and Nasreen Ali. Vaccination against COVID-19: Factors That Influence Vaccine Hesitancy among an Ethnically Diverse Community in the UK. *Vaccines*, 10, 2022.

[24] Dewey Data. Safegraph for academic research. `https://www.deweydata.io/data-partners/safegraph`, 2022. Last Accessed: 2023-08-15.

[25] Sirag Erkol, Dario Mazzilli, and Filippo Radicchi. Effective submodularity of influence maximization on temporal networks. *Physical Review E*, 106, Sep 2022.

[26] Nir Eyal, Anca Gheaus, Axel Gosseries, Monica Magalhaes, Thierry Ngosso, Bastian Steuwer, Viroj Tangcharoensathien, Isa Trifan, and Andrew Williams. Coronavirus Disease 2019 (COVID-19) Vaccine Prioritization in Low- and Middle-Income Countries May Justifiably Depart From High-Income Countries' Age Priorities. *Clinical Infectious Diseases*, Aug 2022.

[27] Shahab Falahi and Azra Kenarkoohi. Host factors and vaccine efficacy: Implications for Covid-19 vaccines. *Journal of Medical Virology*, 94:1330–1335, 2021.

[28] Golnoosh Farnadi, Behrouz Babaki, and Michel Gendreau. A Unifying Framework for Fairness-Aware Influence Maximization. *WWW '20: Companion Proceedings of the Web Conference*, 2020.

[29] Maddalena Ferranna, Daniel Cadarette, and David E Bloom. COVID-19 Vaccine Allocation: Modeling Health Outcomes and Equity Implications of Alternative Strategies. *Engineering (Beijing, China)*, Jul 2021.

[30] Vijay Gayam, Muchi Ditah Chobufo, Mohamed A. Merghani, Shristi Lamichhane, Pavani Reddy Garlapati, and Mark K. Adler. Clinical characteristics and predictors of mortality in African-Americans with COVID-19 from an inner-city community teaching hospital in New York. *Journal of Medical Virology*, 93(2):812–819, Feb 2021.

[31] Nathalie T.H. Gayraud, Evaggelia Pitoura, and Panayiotis Tsaparas. Diffusion Maximization in Evolving Social Networks. *COSN 2015: Proceedings of the 2015 ACM on Conference on Online Social Networks*, 2015.

[32] Amit Goyal, Wei Lu, and Laks V.S. Lakshmanan. CELF++: Optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th International Conference Companion on World Wide Web, WWW*, 2011.

[33] Chinazzi Matteo Gozzi, Nicolò, Natalie E Dean, Ira M Longini, Elizabeth Halloran, Nicola Perra, and Alessandro Vespignani. Estimating the impact of COVID-19 vaccine allocation inequities: a modeling study. *medRxiv*, Nov 2022.

[34] Shasha Han, Jun Cai, Juan Yang, Juanjuan Zhang, Qianhui Wu, Wen Zheng, Huilin Shi, Marco Ajelli, Xiao-Hua Zhou, and Hongjie Yu. Time-varying optimization of COVID-19 vaccine prioritization in the context of limited vaccination capacity. *Nature Communications*, 12, 2021.

[35] Petter Holme. Efficient local strategies for vaccination and network attack. *Europhysics Letters*, 68, 2004.

[36] Heather E. Hsu, Erin M. Ashe, Michael Silverstein, Melissa Hofman, Samantha J. Lange, Hilda Razzaghi, Rebecca G. Mishuris, Ravin Davidoff, Erin M. Parker, Ana Penman-Aguilar, Kristie E.N. Clarke, Anna Goldman, Thea L. James, Karen Jacobson, Karen E. Lasser, Ziming Xuan, Georgina Peacock, Nicole F. Dowling, and Alyson B. Goodman. Race/Ethnicity, Underlying Medical Conditions, Homelessness, and Hospitalization Status of Adult Patients with COVID-19 at an Urban Safety-Net Medical Center - Boston, Massachusetts, 2020. *MMWR. Morbidity and mortality weekly report*, 69(27):864–869, jul 2020.

[37] Peter C. Jentsch, Madhur Anand, and Chris T. Bauch. Prioritising COVID-19 vaccination in changing social and epidemiological landscapes: a mathematical modelling study. *The Lancet Infectious Diseases*, 2021.

[38] Qingye Jiang, Guojie Song, Gao Cong, Yu Wang, Wenjun Si, and Kunqing Xie. Simulated annealing based influence maximization in social networks. *AAAI*, 2011.

[39] Chen Jiangzuo, Stefan Hoops, Achla Marathe, Henning Mortveit, Bryan Lewis, Srinivasan Venkatramanan, Arash Haddadan, Parantapa Bhattacharya, Abhijin Adiga, Anil Vullikanti, Aravind Srinivasan, Mandy Wilson, Gal Ehrlich, Maier Fenster, Stephen Eubank, Christopher Barrett, and Madhav Marathe. Prioritizing allocation of COVID-19 vaccines based on social contacts increases vaccination effectiveness. 2021.

[40] Nicholos P. Joseph, Nicholas J. Reid, Avik Som, Matthew D. Li, Emily P. Hyle, Caitlin M. Dugdale, Min Lang, Joseph R. Betancourt, Francis Deng, Dexter P. Mendoza, Brent P. Little, Anand K. Narayan, and Efrén J. Flores. Racial and Ethnic Disparities in Disease Severity on Admission Chest Radiographs among Patients Admitted with Confirmed Coronavirus Disease 2019: A Retrospective Cohort Study. *Radiology*, 297(3):E303–E312, dec 2020.

[41] Claus Kadelka, Md Rafiul Islam, Audrey McCombs, Jake Alston, and Noah Morton. Ethnic homophily affects vaccine prioritization strategies. *Journal of Theoretical Biology*, Dec 2022.

[42] Amr Kandeel, Ibrahim Eldeyahy, Hanaa Abu ElSood, Manal Fahim, Salma Afifi, Shaimaa Abu Kamar, Hala BahaaEldin, ElSabbah Ahmed, Amira Mohsen, and Khaled Abdelghaffar. COVID-19 vaccination coverage in Egypt: a large-scale national survey – to help achieving vaccination target, March-May, 2022. *BMC Public Health*, 2023.

[43] Jennifer Kates, Jennifer Tolbert, and Josh Michaud. The COVID-19 "Vaccination Line": An Update on State Prioritization Plans. `www.kff.org/coronavirus-covid-19/issue-brief/the-covid-19-vaccination-line-an-update-on-state-prioritization-plans`, Jan 2021. Last Accessed: 2022-10-27.

[44] David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the spread of influence through a social network. *KDD*, 2003.

[45] Moein Khajehnejad, Ahmad Asgharian Rezaei, Mahmoudreza Babaei, Jessica Hoffmann, Mahdi Jalili, and Adrian Weller. Adversarial Graph Embeddings for Fair Influence Maximization over Social Networks. 2020.

[46] Patricia Kipnis, Lauren Soltesz, Gabriel J Escobar, Lauren Myers, and Vincent X Liu. Evaluation of Vaccination Strategies to Compare Efficient and Equitable Vaccine Allocation by Race and Ethnicity Across Time. *JAMA Health Forum*, 2021.

[47] Tony Kirby. Evidence mounts on the disproportionate effect of COVID-19 on ethnic minorities. *The Lancet Respiratory Medicine*, 8, 2020.

[48] Erin Kirwin, Ellen Rafferty, Kate Harback, Jeff Round, and Christopher McCabe. A Net Benefit Approach for the Optimal Allocation of a COVID-19 Vaccine. *PharmacoEconomics*, Sep 2021.

[49] Sungmin Lee, Luis E. C. Rocha, Fredrik Liljeros, and Petter Holme. Exploiting Temporal Network Structures of Human Interaction to Effectively Immunize Populations. *PLOS ONE*, 2012.

[50] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007.

[51] Yuchen Li, Ju Fan, Yanhao Wang, and Kian-Lee Tan. Influence maximization on social graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 30, 2018.

[52] Daniel T. Lichter, Brian C. Thiede, and Matthew M. Brooks. Racial diversity and segregation: Comparing principal cities, inner-ring suburbs, outlying suburbs, and the suburban fringe. *RSF Journal of the Social Sciences*, March 2023.

[53] Andreas Lillebraten, Megan Todd, Jessica Dimka, Nan Zou Bakkeli, and Svenn-Erik Mamelund. Socioeconomic status and disparities in COVID-19 vaccine uptake in Eastern Oslo, Norway. *Public Health in Practice*, 5, 2023.

[54] Dan-Yu Lin, Yu Gu, Bradford Wheeler, Hayley Young, Shannon Holloway, Shadia-Khan Sunny, Zack Moore, and Donglin Zeng. Effectiveness of Covid-19 Vaccines over a 9-Month Period in North Carolina. *The New England Journal of Medicine*, pages 933–941, 2022.

[55] Bo Liu, Gao Cong, Dong Xu, and Yifeng Zeng. Time Constrained Influence Maximization in Social Networks. *IEEE International Conference on Data Mining (ICDM)*, 2012.

[56] Bo Liu, Gao Cong, Yifeng Zeng, Dong Xu, and Yeow Meng Chee. Influence Spreading Path and Its Application to the Time Constrained Social Influence Maximization Problem and Beyond. *IEEE Transactions on Knowledge and Data Engineering*, 26, 2014.

[57] Kaihui Liu and Yijun Lou. Optimizing COVID-19 vaccination programs during vaccine shortages. *Infectious Disease Modelling*, 7, 2022.

[58] Steven Manson, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 17.0 [dataset] Minneapolis, MN: IPUMS. http://doi.org/10.18128/D050.V17.0. 2022.

[59] Laura Matrajt, Julia Eaton, Tiffany Leung, and Elizabeth R. Brown. Vaccine optimization for COVID-19: Who to vaccinate first? *Science Advances*, 7, 2021.

[60] Laura Matrajt, Julia Eaton, Tiffany Leung, Dobromir Dimitrov, Joshua T. Schiffer, David A. Swan, and Holly Janes. Optimizing vaccine allocation for COVID-19 vaccines shows the potential role of single-dose vaccination. *Nature Communications*, 12, 2021.

[61] Zakaria Mehrab, Mandy L. Wilson, Serina Chang, Galen Harrison, Bryan Lewis, Alex Telionis, Justin Crow, Dennis Kim, Scott Spillmann, Kate Peters, Jure Leskovec, and Madhav V. Marathe. Data-Driven Real-Time Strategic Placement of Mobile Vaccine Distribution Sites. *The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*, 2022.

[62] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54:1–35, 2021.

[63] Helmut Meister and Eugen Grycko. Fair Distribution of Vaccines Among Population Groups. 2021.

[64] Gregorio A. Millett, Austin T. Jones, David Benkeser, Stefan Baral, Laina Mercer, Chris Beyrer, Brian Honermann, Elise Lankiewicz, Leandro Mena, Jeffrey S. Crowley, Jennifer Sherwood, and Patrick S. Sullivan. Assessing differential impacts of COVID-19 on black communities. *Annals of epidemiology*, 47:37–44, jul 2020.

[65] Marco Minutoli, Sambaturu Prathyush, Mahantesh Halappanavar, Antonino Tumeo, Ananth Kalyananaraman, and Anil Vullikanti. PREEMPT: Scalable Epidemic Interventions Using Submodular Optimization on Multi-GPU Systems. 2020.

[66] Sabuj Kanti Mistry, Mehrab Ali, Uday Narayan Yadav, Nazmul Huda, Ateeb Ahmad Parray, Rashidul Alam Mahumud, and Dipak Mitra. COVID-19 vaccination coverage is extremely low among older population in Bangladesh: findings from a cross-sectional study. *Human Vaccines Immunotherapeutics*, 18, 2022.

[67] Ramasuri Narayanam and Yadati Narahari. A Shapley Value-Based Approach to Discover Influential Nodes in Social Networks. *IEEE Transaction on Automation Science and Engingeering*, 2011.

[68] Engineering National Academics of Sciences and Medicine. Framework for equitable allocation of Covid-19 vaccine. The National Academies Press, Washington, DC, 2020.

[69] Nambi Ndugga, Latoya Hill, Samantha Artiga, and Sweta Haldar. Latest data on Covid-19 vaccinations by race/ethnicity, KFF. `https://www.kff.org/coronavirus-covid-19/issue-brief/latest-data-on-covid-19-vaccinations-by-race-ethnicity/`, Jul 2022. Last Accessed: 2022-10-26.

[70] Nambi Ndugga, Olivia Pham, Latoya Hill, Artiga Samantha, and Salem Mengistu. Early state vaccination data raise warning flags for racial equity, KFF. `https://www.kff.org/policy-watch/`

early-state-vaccination-data-raise-warning-flags-racial-equity/, Jan 2021. Last Accessed: 2023-10-08.

[71] M.E. J Newman. A measure of betweenness centrality based on random walks. *Social Networks*, pages 39–54, 2005.

[72] Long H Nguyen, David A Drew, Mark S Graham, Amit D Joshi, Chuan-Guo Guo, et al. Risk of COVID-19 among front-line health-care workers and the general community: a prospective cohort study. *The Lancet Public Health*, 5(9):e475–e483, 2020.

[73] The University of Oxford News. 90% of countries have published vaccine plans: Oxford Covid policy tracker. https://www.ox.ac.uk/news/2021-12-09-90-countries-have-published-vaccine-plans-oxford-covid- Last Accessed: 2023-10-08.

[74] Malorie Perry, Ashley Akbari, Simon Cottrell, Michael B. Gravenor, Richard Roberts, Ronan A. Lyons, Stuart Bedston, Fatemah Torabi, and Lucy Griffiths. Inequalities in coverage of COVID-19 vaccination: A population register based cross-sectional study in Wales, UK. *Vaccine*, 39(42):6256–6261, oct 2021.

[75] Eboni G. Price-Haywood, Jeffrey Burton, Daniel Fort, and Leonardo Seoane. Hospitalization and Mortality among Black Patients and White Patients with Covid-19. *The New England journal of medicine*, 382(26):2534–2543, jun 2020.

[76] Zane Rasnača. Essential but unprotected: highly mobile workers in the EU during the COVID-19 pandemic. *ETUI Research Paper-Policy Brief*, 9, 2020.

[77] Cecilia Reyes, Nausheen Husain, Christy Gutowski, Stacy St. Clair, and Gregory Royal Pratt. Chicago's coronavirus disparity: Black Chicagoans are dying at nearly six times the rate of white residents, data show. https://www.chicagotribune.com/coronavirus/ct-coronavirus-chicago-coronavirus-deaths-demographics-lightfoot-20200406-77nlylhiavgjzb2wa4ckivh7m html. Last Accessed: 2023-10-18.

[78] Safegraph. Neighborhood patterns. https://docs.safegraph.com/docs/neighborhood-patterns. Last Accessed: 2023-10-02.

[79] Kazumi Saito, Ryohei Nakano, and Masahiro Kimura. Prediction of Information Diffusion Probabilities for Independent Cascade Model. *Knowledge-Based Intelligent Information and Engineering Systems*, 2008.

[80] Kyle A Sheldrick, Gideon Meyerowitz-Katz, and Greg Tucker-Kellogg. Plausibility of claimed Covid-19 vaccine efficacies by age: A simulation study. *American Journal of Therapeutics*, 2022.

[81] Eunha Shim. Optimal allocation of the limited Covid-19 vaccine supply in South Korea. *JCM: Journal of Clinical Medicine*, 2021.

[82] Don Bambino Geno Tai, Irene G. Sia, Chyke A. Doubeni, and Mark L. Wieland. Disproportionate Impact of COVID-19 on Racial and Ethnic Minority Groups in the United States: a 2021 Update. *Journal of Racial and Ethnic Disparities*, 2022.

[83] Raghavendra Tirupathi, Valeriia Muradova, Raj Shekhar, Sohail Abdul Salim, Jaffar A. Al-Tawfiq, and Venkataraman Palabindala. COVID-19 disparity among racial and ethnic minorities in the US: A cross sectional analysis. *Travel Medicine and Infectious Disease*, 2020.

[84] Yu Wang, Gao Cong, Guojie Song, and Kunqing Xie. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. *KDD*, 2010.

[85] Ruth Elizabeth Watkinson, Richard Williams, Stephanie Gillibrand, Caroline Sanders, and Matt Sutton. Ethnic inequalities in COVID-19 vaccine uptake and comparison to seasonal influenza vaccine uptake in Greater Manchester, UK: A cohort study. *PLOS Medicine*, 2022.

[86] Hao Wu, Kaibo Wang, and Lei Xu. How can age-based vaccine allocation strategies be optimized? A multi-objective optimization framework. *Frontiers in Public Health*, Sep 2022.

[87] Eric Yanchenko, Tsuyoshi Murata, and Petter Holme. Influence maximization on temporal networks: a review. *arXiv preprint arXiv:2307.00181*, 2023.

[88] Yang Ye, Qingpeng Zhang, Xuan Wei, Zhidong Cao, Hsiang-Yu Yuan, and Daniel Dajun Zeng. Equitable access to COVID-19 vaccines makes a life-saving difference to all countries. *Nature Human Behaviour*, 2022.

[89] Ruqaiijah Yearby and Seema Mohaptra. Law, structural racism, and the COVID-19 pandemic. *Journal of Law and Biosciences*, 2020.

[90] Vivian Yee, Simar Singh Bajaj, and Fatima Cody Stanford. Building a pandemic supply chain — equity over equality. *Nature Medicine 2022 28:4*, 28(4):609–610, Mar 2022.

[91] Chuan Zhou, Peng Zhang, Wenyu Zang, and Li Guo. On the upper bounds of spread for greedy algorithms in social network influence maximization. *IEEE Transactions on Knowledge and Data Engineering*, 27, 2015.