On the Cost of Demographic Parity in Influence Maximization

Ruben Becker[1], **Gianlorenzo D'Angelo**[2], **Sajjad Ghobadi**[2]
[1]Ca' Foscari University of Venice, Italy
[2]Gran Sasso Science Institute, L'Aquila, Italy
rubensimon.becker@unive.it, gianlorenzo.dangelo@gssi.it, sajjad.ghobadi@gssi.it

Abstract

Modeling and shaping how information spreads through a network is a major research topic in network analysis. While initially the focus has been mostly on efficiency, recently fairness criteria have been taken into account in this setting. Most work has focused on the maximin criterion however, and thus still different groups can receive very different shares of information. In this work we propose to consider fairness as a notion to be guaranteed by an algorithm rather than as a criterion to be maximized. To this end, we propose three optimization problems that aim at maximizing the overall spread while enforcing strict levels of demographic parity fairness via constraints (either ex-post or ex-ante). The level of fairness hence becomes a user choice rather than a property to be observed upon output. We study this setting from various perspectives. First, we prove that the cost of introducing demographic parity can be high in terms of both overall spread and computational complexity, i.e., the price of fairness may be unbounded for all three problems and optimal solutions are hard to compute, in some case even approximately or when fairness constraints may be violated. For one of our problems, we still design an algorithm with both constant approximation factor and fairness violation. We also give two heuristics that allow the user to choose the tolerated fairness violation. By means of an extensive experimental study, we show that our algorithms perform well in practice, that is, they achieve the best demographic parity fairness values. For certain instances we additionally even obtain an overall spread comparable to the most efficient algorithms that come without any fairness guarantee, indicating that the empirical price of fairness may actually be small when using our algorithms.

Introduction

The internet and particularly online social networks play a central role in how people acquire information nowadays, be it information about political, social, financial, or cultural matters. Several research fields, including mathematics, physics, and computer science, have found interest in analyzing how information spreads through networks. Besides abstractions to (probabilistically) model information spread, the main contributions of computer science in this context have been algorithmic ones. Among them, probably most importantly, the question on how to spread information *efficiently* through a network. More precisely, given a social network and a probabilistic model on how information propagates through it, the main addressed question has been the following: Which *seed set* of size at most k (an input parameter) to target such that the expected number of nodes that obtain the information is maximized, when the information spreads from the chosen seed set? This problem, called *influence maximization*, has received a lot of attention by computer science researchers in diverse communities, including algorithms (e.g. (Kempe, Kleinberg, and Tardos 2015; Borgs et al. 2014; Sadeh, Cohen, and Kaplan 2020)), artificial intelligence (e.g. (Wilder et al. 2018a; Yadav et al. 2018; Becker et al. 2020)), and data and graph mining (e.g. (Cohen et al. 2014; Tang, Xiao, and Shi 2014; Tang, Shi, and Xiao 2015; Chen and Teng 2017; Wu et al. 2019)). As a result the problem is well understood from many perspectives, among them theoretical complexity, approximation algorithms, adaptivity, and practically efficient implementations.

As access to information via social networks may have a big impact on our life, see, e.g. (Banerjee et al. 2013), researchers have taken also *fairness* issues with respect to information spread into account, see the related work below for a non-exhaustive list. In these works, the social network is composed of individuals or groups of individuals (called communities) and the goal is to provide similar information access to all of them. In other words, the focus is not restricted to the efficiency of the information spread, but rather on assuring that each of the communities gets its fair share of information (or coverage). Here, an essential question arises, namely: What do we mean by fair? There is a large variety of fairness notions (Barocas, Hardt, and Narayanan 2019) and in fact different notions have been investigated also in this scope, with the most common one being the maximin criterion (Tsang et al. 2019; Fish et al. 2019; Becker et al. 2022). Here, the goal is to develop algorithms that maximize the minimum coverage of any community or individual (the special case of singleton communities). In some works, where the focus is on communities, this notion is also referred to as *group fairness* or *demographic parity*. What all three previously mentioned works, have in common however is that they consider fairness as a measure to be optimized, namely via maximizing the minimum coverage.

This raises, however, a conceptual question. When maximizing the minimum coverage, we may still end up in a situation where the values of two groups differ a lot. More precisely, consider an example with two groups, say C and D. All the three mentioned approaches would prefer an outcome where C gets a coverage of 0.5 while D gets a coverage of 1 over an outcome where both receive a coverage of 0.499. Now, while fairness is a debatable concept, the second outcome may be considered more fair by many. In fact, if we take a closer look at what is typically understood under group fairness or demographic parity, for example in the machine learning community, see, e.g., Definition 1 in Chapter 2 in the book by Barocas, Hardt, and Narayanan (2019), we observe that, demographic parity (also referred to as independence) is actually defined as *equality* in probability of being selected conditioned on group membership. In the above example, this is satisfied in the second outcome, but far from being satisfied in the first. More fundamentally, the following question arises. In all of these works fairness is considered as a notion to be optimized. But is this the right way of considering fairness? Is fairness not instead something that we want algorithms to *guarantee*, i.e., don't we want to restrict algorithms to satisfy certain levels of fairness independent of their objective?

Our Contribution. In this work, we adopt a different and more strict view on fairness, that is, we consider fairness as a requirement that has to be ensured by the algorithm rather than a notion to be maximized. In terms of the optimization problems at hand, this results in fairness being taken into account via constraints instead of in the objective function, the obvious advantage being that the resulting fairness violation is strictly bounded. More precisely, we develop optimization problems that aim to maximize the overall spread (or coverage) while ensuring that the coverage of all groups is identical, in this way enforcing demographic parity.

While such a strict fairness notion may easily result in infeasibility, we show how to bypass this problem by using an approach popular in economics and computational social choice: we study also *ex-ante fairness* rather than just ex-post fairness. This approach, that was first used in the context of influence maximization by Becker et al. (2022), allows probabilistic rather than deterministic solutions, i.e., distributions over seed sets instead of single sets. Then the expected group coverage when a set is sampled according to this distribution is considered instead of simply the group coverage of a group from a single seed set. This approach is not only useful for the purpose of feasibility, but instead offers various advantages, see, e.g., (Brandl, Brandt, and Seedig 2016; Aziz, Brandt, and Stursberg 2013; Bogomolnaia and Moulin 2001; Katta and Sethuraman 2006). See also the illustrative example of Machina (1989), where a parent assigns an (indivisible) treat to one of two children.

It is clear that such a strict approach to fairness as adopted here may lead to a big loss in efficiency, i.e., in overall spread and possibly also in time complexity of respective algorithms. One of our contributions, is to rigorously analyze these two kinds of loss. We in fact prove in Section Influence Maximization under Demographic Parity that the *price of fairness* may be unbounded in this context. We then proceed by studying the complexity of the proposed optimization problems, more precisely their approximation properties. This includes both proving hardness of approximation results, see Section Hardness Results, and developing an approximation algorithm, see Section Algorithms for $\mathrm{iIM}^{\mathrm{dp}}$ and $\mathrm{pIM}^{\mathrm{dp}}$. Due to space limitations, all proofs are deferred to the full version (Becker, D'Angelo, and Ghobadi 2023). Our study here explicitly includes bi-criteria approximation, that is, we relax the fairness constraints or allow them to be violated within a limited amount (multiplicatively or additively). This permits us to propose algorithms that entitle the user to choose the tolerated amount of fairness violation freely instead of observing the fairness violation upon seeing the output of the algorithm. We proceed by developing efficient heuristics for the proposed problems and conclude with a detailed experimental study on the performance of the developed algorithms both in terms of efficiency and fairness in Section Experiments. For our experiments, we use random, synthetic, and real-world data sets. Our experimental study shows that although our theoretical results are mainly pessimistic, our algorithms achieve a trade-off between fairness and overall coverage and in some cases even achieve similar coverage as state-of-the-art influence maximization algorithms while guaranteeing fairness on top.

Related Work. Fish et al. (2019) were the first to study the maximin criterion in influence maximization, they focus on individual fairness and show NP-hardness as well as that the problem is hard to approximate unless $P = NP$. Tsang et al. (2019) study the maximin criterion and diversity constraints with respect to groups. They give an algorithm with asymptotic approximation factor $1-1/e$ in the setting where there are $o(k \log^3 k)$ communities. The work that is probably closest to ours is the one by Becker et al. (2022). Also this work uses the maximin criterion for group fairness, rather than demographic parity in the exact sense of its definition. Still, similar to ours, this work allows probabilistic seeding strategies.

Stoica and Chaintreau (2019) define "fairness in outreach" that is essentially equivalent to demographic parity. Their work however does not introduce tailored algorithms but is instead more focused on analyzing the fairness achieved by standard algorithms for influence maximization. Farnadi, Barbaki, and Gendreau (2020) propose a framework for fair influence maximization that is based on mixed integer linear programs (MILPs). Their framework, that is unlikely to be applicable to large instances, captures various notions of group fairness, including "equity", which again coincides with demographic parity. Ali et al. (2022) address fairness in influence maximization within a time-critical setting. The authors also consider fairness notions that are similar to demographic parity, but instead of maintaining the fairness constraints, they pass the group coverages through some monotone concave function and include it in the objective. Other somewhat related works include (Gershtein, Milo, and Youngmann 2021; Stoica, Han, and Chaintreau 2020; Anwar, Saveski, and Roy 2021; Rahmattalabi et al. 2021; Khajehnejad et al. 2020; Wang, Varol, and Eliassi-Rad 2021).

Preliminaries

Information Diffusion. In the classical influence maximization setting, we are given a directed graph $G = (V, E)$ with $|V| = n$ and edge weights $\{w_e \in [0,1] : e \in E\}$. We use the *Triggering model* (Kempe, Kleinberg, and Tardos 2015) for describing the random process of information diffusion. The Triggering model is a generalization of both the *Independent Cascade (IC)* and *Linear Threshold (LT) models*. For a *seed set* $A \subseteq V$, the spread $\sigma(A)$ from A is the expected number of nodes reached from A in a random sample of *triggering sets* which is constructed as follows. Every node $v \in V$ independently picks a *triggering set* T_v among its in-neighbors N_v according to some distribution. Let $L = (T_v)_{v \in V}$ be a possible outcome of sampled triggering sets; L defines a *live-edge graph* $G_L = (V, E_L)$, where $E_L = \{(u,v) | v \in V, u \in T_v\}$. Then $\rho_L(A)$ is the set of nodes reachable from A in G_L and the *expected spread of* A is $\sigma(A) := \mathbb{E}_{\mathcal{L}}[|\rho_{\mathcal{L}}(A)|]$, where $\mathcal{L}$ denotes a random live-edge graph. We also use the term *overall coverage* for the expected fraction of reached nodes $\sigma(A)/|V|$. We obtain the IC model from the Triggering model if, for each edge (u, v), the node u is added to the T_v with probability w_{uv}. Differently, in the LT model each v picks at most one of its in-neighbors u with probability $w_{(u,v)}$.

Approximation Algorithms. For $N \in \mathbb{N}$, we use $[N]$ to denote the integers from 1 to N. We will consider maximization problems of the form $\max\{F(x) : x \in R \text{ and } \exists\gamma : A_i(x) = \gamma \text{ for all } i \in [m]\}$, where R is a feasibility region, the functions $A_i : R \to \mathbb{R}_{\geq 0}$, for $i \in [m]$, define a set of (additional) constraints, and $F : R \to \mathbb{R}_{\geq 0}$ is an objective function. We consider approximation algorithms (possibly) with constraint violation. Let $\alpha, \beta \in (0, 1]$ be real values. Then, we say that $x \in R$ is *β-feasible* if $A_i(x) \geq \beta A_j(x)$ for all pairs of $i, j \in [m]$. We say that $x \in R$ is an *(α, β)-approximation* if x is β-feasible and $F(x) \geq \alpha \, \mathrm{opt}$, where opt is the optimum value. We call an algorithm a *(α, β)-approximation algorithm*, if it is a polynomial-time algorithm whose output solutions are (α, β)-approximations.

Influence Maximization under Demographic Parity

In the classical influence maximization problem (IM), given a graph G and an integer k, the objective is to find a set of k seeds that maximizes the expected spread, i.e., $\max_{S \in \mathcal{S}}\{\sigma(S)\}$, where $\mathcal{S} := \{S \subseteq V : |S| \leq k\}$ is the set of subsets of nodes of size at most k. We refer to the optimal value of this optimization problem as $\mathrm{opt}(G, k)$.

Requiring Demographic Parity. In our setting, in addition to G and k, we are given a *community structure* $\mathcal{C}$ that is a set of m non-empty communities $C \subseteq V$. Notice that communities may neither be disjoint nor cover the whole node set. Our goal now is to find a set S of size at most k that maximizes the total spread while the fraction of reached nodes in each community is the same among all communities, i.e., achieving perfect demographic parity. To make this formal, we introduce $\sigma_v(S) := \Pr_{\mathcal{L}}[v \in \rho_{\mathcal{L}}(S)]$ as the probability that node v is reached from S. Note that the expected spread is the sum over all these probabilities, i.e., $\sigma(S) = \mathbb{E}_{\mathcal{L}}[|\rho_{\mathcal{L}}(S)|] = \sum_{v \in V} \sigma_v(S)$. For a community $C \in \mathcal{C}$, we then denote by $\sigma_C(S) := \frac{1}{|C|} \cdot \sum_{v \in C} \sigma_v(S)$ the average probability of nodes being reached in C or equivalently this is the *expected group coverage of* $\mathcal{C}$, i.e., the expected fraction of nodes from $\mathcal{C}$ that are reached. We are now ready to formally define our first optimization problem, we refer to it as $\mathrm{IM}^{\mathrm{dp}}$, standing for influence maximization under demographic parity:

$$\max_{S \in \mathcal{S}} \left\{\sigma(S) : \exists\gamma : \sigma_C(S) = \gamma \text{ for all } C \in \mathcal{C}\right\}. \quad (\mathrm{IM}^{\mathrm{dp}})$$

For an instance, consisting of a graph G, communities $\mathcal{C}$, and an integer k, we call $\mathrm{opt}_{\mathcal{S}}(G, \mathcal{C}, k)$ the optimum of $\mathrm{IM}^{\mathrm{dp}}$.

Fairness via Randomization. In addition to $\mathrm{IM}^{\mathrm{dp}}$, we define optimization problems that permit randomized strategies in the seed selection process rather than only deterministic ones, in an analogous way to what Becker et al. (2022) did for the maximin criterion. Inspired by Becker et al., we introduce two different probabilistic settings, a general one and one that chooses seed nodes independently.

In the first problem, $\mathrm{pIM}^{\mathrm{dp}}$, standing for probabilistic influence maximization under demographic parity, feasible solutions are distributions over node sets. Formally, we let $\mathcal{P} := \{p \in [0,1]^{2^V} : \mathbb{1}^T p = 1, \sum_{S \subseteq V} p_S |S| \leq k\}$ be the set of distributions over node sets of expected size at most k and denote by $S \sim p$ the random process of sampling S according to $p \in \mathcal{P}$. Now, the goal in $\mathrm{pIM}^{\mathrm{dp}}$ is to find the distribution $p \in \mathcal{P}$ that maximizes the expected number of reached nodes, while ensuring that perfect demographic parity is satisfied in expectation, i.e., that the expected probability to be reached is the same among all communities. Formally, $\mathrm{pIM}^{\mathrm{dp}}$ is defined as

$$\max_{p \in \mathcal{P}} \{\sigma(p) : \exists\gamma \text{ s.t. } \sigma_C(p) = \gamma \text{ for all } C \in \mathcal{C}\}, \quad (\mathrm{pIM}^{\mathrm{dp}})$$

where we extend set functions to vectors in a straightforward way, i.e., for a set function f, we let $f(p) := \mathbb{E}_{S \sim p}[f(S)]$. For an instance $G, \mathcal{C}, k$, $\mathrm{opt}_{\mathcal{P}}(G, \mathcal{C}, k)$ is the optimum.

In the second probabilistic variant of $\mathrm{IM}^{\mathrm{dp}}$, we restrict to independent probability distributions, that is, in a feasible solution each node is selected as a seed independently with some probability in such a way that the expected size of the seed set is at most k. Formally, we let $\mathcal{I} := \{x \in [0,1]^n : \mathbb{1}^T x \leq k\}$ and, for $x \in \mathcal{I}$, we denote with $S \sim x$ the process of randomly generating a set S from x, where each i is included in S independently with probability x_i. We then obtain independent probabilistic influence maximization under demographic parity problem $\mathrm{iIM}^{\mathrm{dp}}$ as:

$$\max_{x \in \mathcal{I}} \{\sigma(x) : \exists\gamma \text{ s.t. } \sigma_C(x) = \gamma \text{ for all } C \in \mathcal{C}\}, \quad (\mathrm{iIM}^{\mathrm{dp}})$$

where again for a set function f and a vector $x \in \mathcal{I}$, we let $f(x) := \mathbb{E}_{S \sim x}[f(S)]$. Again, for an instance $G, \mathcal{C}, k$, we denote with $\mathrm{opt}_{\mathcal{I}}(G, \mathcal{C}, k)$ the optimum of $\mathrm{iIM}^{\mathrm{dp}}$.

Finally, we note that Becker et al. (2022) refer to the two variants of the above problems in their setting of the maximin criterion as set-based and node-based problem.

Demographic Parity vs. Maximin. We proceed by giving an example that illustrates that considering the maximin criterion as done by Becker et al. and demographic parity in our strict sense can lead to drastically different outcomes. More precisely, we construct an instance where the optimal maximin solution suffers linear multiplicative violation in demographic parity, while achieving an expected coverage that is only around twice as good as a solution that achieves perfect demographic parity. This is formalized below.

Lemma 0.1. *Let $\varepsilon > 0$. There is an instance $G, \mathcal{C}, k$ with n nodes, in which the optimal maximin strategy achieves an overall expected coverage of $2+\varepsilon$, but suffers a violation in demographic parity of $(n-1)/(1+\varepsilon) = \Theta(n)$. On the other hand,* $\mathrm{opt}_{\mathcal{P}}(G, \mathcal{C}, k) = (n+1)/(n-\varepsilon) = 1 + \Theta(1/n)$.

Relationship between $\mathrm{IM}^{\mathrm{dp}}$, $\mathrm{pIM}^{\mathrm{dp}}$, and $\mathrm{iIM}^{\mathrm{dp}}$. We first observe that clearly every feasible solution of $\mathrm{IM}^{\mathrm{dp}}$ corresponds to a feasible solution of $\mathrm{iIM}^{\mathrm{dp}}$ and $\mathrm{pIM}^{\mathrm{dp}}$, respectively. Furthermore, every feasible solution of $\mathrm{iIM}^{\mathrm{dp}}$ directly corresponds to a feasible solution of $\mathrm{pIM}^{\mathrm{dp}}$ via the following transformation: For $x \in \mathcal{I}$ define the vector p^x as $p^x_S := \prod_{i\in S} x_i \prod_{j\in V\setminus S}(1-x_j)$, for $S \subseteq V$. Then, observe that $\sigma(x) = \sigma(p^x)$, $p^x \in \mathcal{P}$, and $\sigma_C(x) = \sigma_C(p^x)$, for any $C \in \mathcal{C}$. Hence, we obtain the following lemma.

Lemma 0.2. *For every instance $G, \mathcal{C}, k$, it holds that*

$$\mathrm{opt}_{\mathcal{S}}(G, \mathcal{C}, k) \le \mathrm{opt}_{\mathcal{I}}(G, \mathcal{C}, k) \le \mathrm{opt}_{\mathcal{P}}(G, \mathcal{C}, k).$$

A natural question is then whether a similar relation holds also in the other direction. We observe that this is not the case, $\mathrm{opt}_{\mathcal{I}}(G, \mathcal{C}, k)$ cannot be upper bounded in terms of $\mathrm{opt}_{\mathcal{S}}(G, \mathcal{C}, k)$ multiplicatively and $\mathrm{opt}_{\mathcal{P}}(G, \mathcal{C}, k)$ not in terms of $\mathrm{opt}_{\mathcal{I}}(G, \mathcal{C}, k)$. Formally:

Lemma 0.3. *Assume information spread to follow the IC model. There exist instances $G, \mathcal{C}, k$ s.t.*

$$(i)\ \frac{\mathrm{opt}_{\mathcal{S}}(G, \mathcal{C}, k)}{\mathrm{opt}_{\mathcal{I}}(G, \mathcal{C}, k)} = 0, \textit{ and } (ii)\ \frac{\mathrm{opt}_{\mathcal{I}}(G, \mathcal{C}, k)}{\mathrm{opt}_{\mathcal{P}}(G, \mathcal{C}, k)} = 0$$

as well as $(iii)\ \mathrm{opt}_{\mathcal{P}}(G, \mathcal{C}, k) - \mathrm{opt}_{\mathcal{I}}(G, \mathcal{C}, k) = \Omega(n)$.

Price of Fairness. The price of (group) fairness is a measure of loss in efficiency due to fairness. More precisely, for $X \in \{\mathcal{S}, \mathcal{I}, \mathcal{P}\}$, we define $\mathrm{PoF}_X(G, \mathcal{C}, k)$ as the ratio of the maximum coverage in the absence of fairness constraints, i.e., $\mathrm{opt}(G, k)$ to the optima of the corresponding problem involving demographic parity fairness constraints, in other words, $\mathrm{PoF}_X(G, \mathcal{C}, k) := \mathrm{opt}(G, k)/\mathrm{opt}_X(G, \mathcal{C}, k)$. Due to Lemma 0.2, we have the following relation $\mathrm{PoF}_{\mathcal{S}}(G, \mathcal{C}, k) \ge \mathrm{PoF}_{\mathcal{I}}(G, \mathcal{C}, k) \ge \mathrm{PoF}_{\mathcal{P}}(G, \mathcal{C}, k)$. We proceed by showing that the PoF can be unbounded for $\mathrm{pIM}^{\mathrm{dp}}$ and thus in all three cases.

Lemma 0.4. *Assume that information spread follows the IC model. For any even $n > 0$, there is an instance $G, \mathcal{C}, k$ s.t.* $\mathrm{PoF}_X(G, \mathcal{C}, k) = \Omega(n)$ *for* $X \in \{\mathcal{S}, \mathcal{I}, \mathcal{P}\}$.

Hardness Results

In this section, we give several hardness and hardness of approximation results for $\mathrm{IM}^{\mathrm{dp}}$, $\mathrm{pIM}^{\mathrm{dp}}$, and $\mathrm{iIM}^{\mathrm{dp}}$.

Hardness of $\mathrm{IM}^{\mathrm{dp}}$. We first show that it is NP-hard to approximate $\mathrm{IM}^{\mathrm{dp}}$ to within any bounded factor. Indeed, we prove two stronger and more general statements: One cannot find in polynomial time a solution that approximates the optimum of $\mathrm{IM}^{\mathrm{dp}}$, even if we allow the fairness constraints to be violated by a multiplicative or an additive term, unless $P = NP$. We start with the multiplicative case.

Theorem 0.5. *For any $\alpha \in (0, 1]$, $\beta \in (0, 1]$, there is no (α, β)-approximation algorithm for $\mathrm{IM}^{\mathrm{dp}}$, unless $P = NP$.*

The proof is via a reduction from SETCOVER. We now turn to the additive case. For a given $\varepsilon \in [0, 1)$, we say that a seed set S is ε^+-feasible if $\left|\sigma_{C_i}(S) - \sigma_{C_j}(S)\right| \le \varepsilon$ for all $C_i, C_j \in \mathcal{C}$, $i \ne j$. For $\alpha \in (0, 1]$ and $\varepsilon \in [0, 1)$, an $(\alpha, \varepsilon)^+$-approximation algorithm for $\mathrm{IM}^{\mathrm{dp}}$ produces an ε^+-feasible seed set S such that $\sigma(S) \ge \alpha\,\mathrm{opt}$. Using a similar reduction we show the following theorem.

Theorem 0.6. *For $\alpha \in (0, 1]$, $\varepsilon \in [0, 1)$, there is no $(\alpha, \varepsilon)^+$-approximation algorithm for $\mathrm{IM}^{\mathrm{dp}}$, unless $P = NP$.*

Hardness of $\mathrm{pIM}^{\mathrm{dp}}$ and $\mathrm{iIM}^{\mathrm{dp}}$. We prove the following theorem, again via a reduction from SETCOVER.

Theorem 0.7. *The $\mathrm{pIM}^{\mathrm{dp}}$ problem is NP-hard.*

For $\mathrm{iIM}^{\mathrm{dp}}$ we show an ever stronger result via a reduction from MAXCOVERAGE: It cannot be approximated better than within $1 - 1/e$, unless $P = NP$.

Theorem 0.8. *There is no $(\alpha, 0)$-approximation algorithm for $\mathrm{iIM}^{\mathrm{dp}}$ for a constant $\alpha > 1 - 1/e$, unless $P = NP$.*

Algorithms for $\mathrm{iIM}^{\mathrm{dp}}$ and $\mathrm{pIM}^{\mathrm{dp}}$

We proceed with algorithms for $\mathrm{iIM}^{\mathrm{dp}}$ and $\mathrm{pIM}^{\mathrm{dp}}$. First note that it is not feasible to evaluate the functions σ and σ_C involved in the optimization problems exactly. It is however well understood that the functions can be approximated using sampling. For the following discussion, we simply assume access to approximations $\tilde{\sigma}$ and $\tilde{\sigma}_C$.

Approximation Algorithm for $\mathrm{iIM}^{\mathrm{dp}}$. We start by giving an approximation algorithm for $\mathrm{iIM}^{\mathrm{dp}}$. Given the above discussion, we consider $\tilde{\sigma}$ and $\tilde{\sigma}_C$ instead of σ and σ_C:

$$\max_{x\in\mathcal{I}}\{\tilde{\sigma}(x) : \exists\gamma \text{ s.t. } \tilde{\sigma}_C(x) = \gamma\ \forall C \in \mathcal{C}\}. \quad \text{apx-ipIM}^{\mathrm{dp}}$$

As discussed above, an (α, β)-approximation x for an instance $(G, \mathcal{C}, k)$ of $\text{apx-ipIM}^{\mathrm{dp}}$ approximates $\mathrm{iIM}^{\mathrm{dp}}$ by adding a multiplicative error in the objective and an additive error in the fairness violation, that is it satisfies $\sigma(x) \ge (\alpha - \varepsilon)\,\mathrm{opt}_{\mathcal{I}}(G, \mathcal{C}, k)$ and $\sigma_C(x) \ge \beta\sigma_{C'}(x) - \varepsilon$, for any arbitrary small $\varepsilon > 0$. We can thus focus on giving an approximation algorithm for $\text{apx-ipIM}^{\mathrm{dp}}$. Formally, we prove:

Theorem 0.9. *There exists a $(1 - 1/e, 1 - 1/e)$-approximation algorithm for* $\text{apx-ipIM}^{\mathrm{dp}}$.

We first note that the objective function of $\text{apx-ipIM}^{\mathrm{dp}}$ is not linear, since the probability of sampling a seed set S from a distribution $x \in \mathcal{I}$ is $\prod_{i\in S} x_i \prod_{i\notin S}(1 - x_i)$. Our approach here is to approximate $\text{apx-ipIM}^{\mathrm{dp}}$ by a linear program (LP) of polynomial size. In our experimental study we refer to the described algorithm as **`ind_lp`**.

Algorithms for pIM^{dp}. In this paragraph, we present algorithms for pIM^{dp} that are based on greedy strategies and solving a (comparatively) small linear program. We again focus on the problem with the approximate functions $\tilde{\sigma}$ and $\tilde{\sigma}_C$ and refer to it as apx-pIM^{dp} (it is defined analogously to apx-ipIM^{dp}). Differently from apx-ipIM^{dp}, the objective function of apx-pIM^{dp} is linear and hence it can be formulated as a linear program by introducing a variable for each seed set $S \subseteq 2^V$. However, the size of such a linear program would be $\Theta(2^n)$, the dimension of $\mathcal{P}$. Our approach here is to restrict to a subset $\mathcal{Q} \subseteq \mathcal{P}$ in such a way that the linear program at hand becomes more tractable. More precisely, the two heuristics that we propose are based on solving the following linear program for two different choices of $\mathcal{Q}$

$$\max_{p\in\mathcal{Q}}\{\tilde{\sigma}(p) : \exists\gamma \text{ s.t. } \tilde{\sigma}_C(p) = \gamma \text{ for all } C \in \mathcal{C}\}. \qquad \text{P}_{\mathcal{Q}}$$

In the first heuristic, **grdy_grp+lp**, we choose $\mathcal{Q}$ by restricting the set of non-zero variables to sets that either (1) have a large coverage with respect to a certain community, or (2) have a large overall coverage. Formally, $\mathcal{Q} := \{p \in \mathcal{P} : p_S = 0 \text{ for all } S \notin \mathcal{S}_1 \cup \mathcal{S}_2\}$, where $\mathcal{S}_1 = \{S_i : i \in [m]\}$ with $S_i = \text{argmax}_{S\in V}\{\tilde{\sigma}_{C_i}(S) : |S| \leq k\}$, $\mathcal{S}_2 := \{T_i : i \in \{0\} \cup [2k]\}$ with $T_i := \text{argmax}_{S\in\mathcal{S}}\{\tilde{\sigma}(S) : |S| \leq i\}$. Here the choice of $2k$ in the definition of $\mathcal{S}_2$ is more or less arbitrary, the rationale being that due to submodularity of σ it is unlikely that choosing a set of size twice the allowed expectation leads to a profitable gain in overall spread. Clearly, the idea behind this choice of $\mathcal{Q}$ is to provide the LP with sufficiently many degrees of freedom to both achieve a high overall coverage and a good coverage for each community.

In the second heuristic, **maxmin+lp**, we define $\mathcal{Q} := \{p \in \mathcal{P} : p = \lambda_0 \cdot \mathbb{1}_\emptyset + \sum_{i\in[m]} \lambda_i \mathbb{1}_{S_i} + \lambda_{m+1} q\}$, where $\mathbb{1}_S$ is the 2^n-dimensional vector that is 1 at position $S \subseteq V$ and zero elsewhere, and $q \in \mathcal{P}$ is the distribution computed by the algorithm of Becker et al. (2022) for the maximin criterion. In other words, we restrict to probability distributions in $\mathcal{P}$ that are linear combinations of (1) a distribution computed for the maximin criterion and (2) the degenerate distributions of the empty set and the sets maximizing the respective community coverage. The rationale of this choice of $\mathcal{Q}$ is to profit from the efficiency of the maximin solution but enabling the LP solver to improve the incurred violation in demographic parity by putting additional probability on the deterministic distributions corresponding to under-represented communities.

Experiments

In this section, we report on a detailed experimental study. We evaluate a diverse set of algorithms for influence maximization in terms of their efficiency (both overall coverage and run-time) and demographic parity fairness. [1] In our evaluation, we use random, synthetic, and real data sets.

Algorithms. In addition to **ind_lp**, **grdy_grp+lp**, and **maxmin+lp**, our study includes the following competitors:

grdy_im the greedy algorithm for IM,

grdy_maxmin the algorithm that greedily maximizes the minimum community coverage,

grdy_prop a simple heuristic that greedily maximizes σ_{C_i} for $i \in [m]$ using $k|C_i|/n$ seeds,

milp the MILP of Farnadi, Babaki, and Gendreau (2020),

moso an algorithm based on multi-objective submodular optimization due to Tsang et al. (2019),

mult_weight the multiplicative weights routine for the set-based problem of Becker et al. (2022),

myopic a simple heuristic by Fish et al. (2019), and

uniform the uniform solution to iIM^{dp}.

We refer to the original papers for details about **moso** and **mult_weight**.

The **myopic** heuristic, after choosing the node of maximum degree in the first iteration, always selects the node with minimum probability of being reached. We note that **grdy_maxmin**, **mult_weight**, **moso**, and **myopic** were designed for the maximin criterion. We emphasize that **mult_weight**, **ind_lp**, **grdy_grp+lp**, **maxmin+lp**, and **uniform** compute distributions and are thus designed for achieving ex-ante guarantees, while the other algorithms compute deterministic seed sets. For our algorithms from the previous section we relax the strict demographic parity constraints for some parameter $\eta \in [0, 1)$ as follows. For **grdy_grp+lp** and **maxmin+lp**, we replace γ in the demographic parity constraints in $\text{P}_{\mathcal{Q}}$ by $\gamma \pm \eta$ for $\eta \in \{0, x/16, x/8, x/4\}$, where x is the violation in demographic parity that **grdy_im** suffers. For **ind_lp** we do a similar relaxation.

Instances. We use random, synthetic and real world graphs. (1) Our random graphs are generated using the Barabasi-Albert model with parameter $m = 2$, i.e., connecting a newly added node to two existing nodes. (2) The synthetic networks are the ones used by Tsang et al. (2019) that go back to the work of Wilder et al. (2018b). Every node in these networks is associated with some attributes (region, ethnicity, age, gender and status) and nodes with the same attributes are more likely to connect to each other. Each network consists of 500 nodes and the attributes induce communities. (3) We use the same set of real world instances as Fish et al. (2019). We considered the largest weakly connected component for all these graphs in order to make fair coverage more achievable. We use the IC model with uniformly random weights in $[0, 0.4]$ for the random and synthetic networks and $[0, 0.2]$ for real world instances.

We consider the following different community structures. (1) Singleton communities: each node forms its own community. (2) Random communities: each node is assigned u.a.r. to a community. (3) BFS communities: for a predefined number of communities m, each community of size n/m is generated by a breadth first search from a random source node (if the size of community does not reach n/m, we pick a new random node and continue the process), this results in rather connected communities. (4) Random-overlap communities: for a given m, a node is, each with probability $1/(m + 2)$, (i) in community C_i for $i \in [m]$,

[1] https://github.com/sajjad-ghobadi/demographic_parity.git

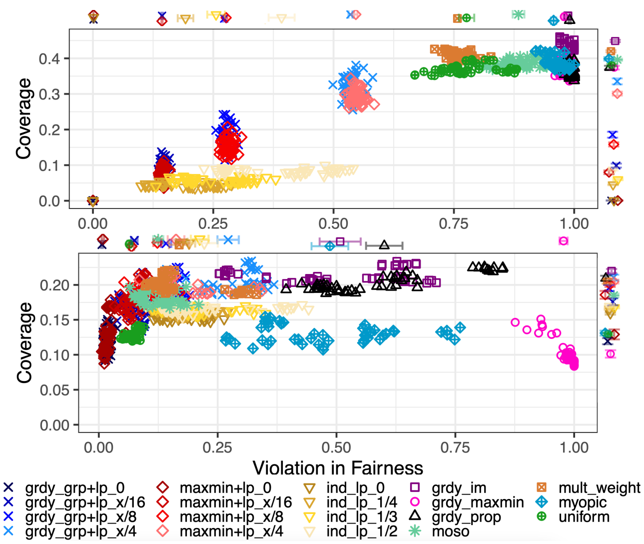

Figure 1: (1) Random instances ($k = 25$, $n = 200$, singleton communities), (2) synthetic instances ($k = 25$, $n = 500$, communities induced by gender and region).

(ii) in no community, or (iii) in all m communities. (5) Leidenalg communities: communities detected by a common algorithm for community detection (Traag, Waltman, and van Eck 2019). (6) Given communities for the synthetic networks and for some of the real world instances.

Experimental Setting. As all evaluated algorithms are randomized, we repeat each run 10 times per graph, for random and synthetic graphs, we in addition average over 5 graphs, thus resulting in 50 runs per algorithm. In all our 2-dimensional plots, we also show averages of the projections onto each dimension together with 95% confidence intervals. For algorithms that output distributions rather than sets, i.e., giving ex-ante guarantees, we evaluate both their overall coverage and their demographic parity violation *in expectation*. We ran experiments with a large variety of parameter settings and, due to space limitations, can only report on a subset of the experiments performed. In our plots the overall (expected) coverage (as ratio of overall nodes) is on the vertical axis while the violation in demographic parity is on the horizontal axis. We note that a perfect algorithm would achieve maximum overall coverage, while suffering zero violation in demographic parity, thus ending up in the top left of the plots.

Running Times. We measure the running times of all algorithms on the random instances for increasing values of $n = 50, 100, 200$. We exclude `uniform` as it takes constant time and `milp` for $n > 50$ as it does not terminate in less than 30 mins. `grdy_im`, `ind_lp`, and `myopic` are fastest. As we will see, unfortunately, the fairness achieved by `grdy_im` and `myopic` is very poor. From the competitor algorithms, `grdy_maxmin`, `milp` and `moso` perform the worst in terms of running times and as their fairness values are not too good either, we exclude them from experiments involving the real-world instances.

Results for Random and Synthetic Networks. We start with the random networks, see the top of Figure 1. We exclude `milp` from this and all further experiment as it does not solve a single instance in less than 30 mins. All competitor algorithms suffer a fairness violation of more than 0.75 and achieve a coverage between 0.35 and 0.45. In the case of `grdy_im`, there is a fairness violation of almost 1. Next, note that our algorithms that are restricted to find perfectly fair solutions, i.e., `grdy_grp+lp_0`, `maxmin+lp_0`, and `ind_lp_0` obtain zero overall coverage. As we are in the setting of singleton communities, perfect demographic parity is a very strong requirement. Instead, if we use `grdy_grp+lp_x/4` (`maxmin+lp_x/4`), where x is the violation of `grdy_im` (here ≈ 1), we still achieve 75% (67%) of `grdy_im`'s coverage while suffering a fairness violation of only 0.5. More generally, `grdy_grp+lp` and `maxmin+lp` allow for a trade-off between coverage and fairness. If the user is for example willing to tolerate only a fairness violation of around 0.25, he can use `grdy_grp+lp_x/8` (or `maxmin+lp_x/8`) and would still achieve 41% (or 35%) of `grdy_im`'s coverage. Note that the algorithm `ind_lp` performs worse than `grdy_grp+lp` and `maxmin+lp` in terms of coverage with similar fairness values.

For the synthetic data sets of Wilder et al. (2018b), see the lower plot in Figure 1, we show results for the community structure induced by the attributes gender and region consisting of 15 communities of largely varying sizes. The best competitor algorithm in terms of fairness violation is `uniform` with a fairness violation of around 0.07, on the other hand it achieves a coverage of only around 0.13. The `moso` algorithm of Tsang et al. (2019) achieves a fairness violation of around 0.13 while achieving a coverage of around 0.18. The `grdy_im` algorithm achieves the biggest coverage of around 0.21, but suffers a huge fairness violation of around 0.5. Here, our algorithms `grdy_grp+lp` and `maxmin+lp` even achieve a decent overall coverage of 55% and 60% of `grdy_im`'s (comparable to, e.g., `moso`) when we restrict to no fairness violation at all (note that there is still a tiny violation in fairness as the final evaluation is done with an independent sample of live-edge graphs). Furthermore, when we allow a fairness violation of $x/16$, where x is the violation of `grdy_im`, our algorithms `grdy_grp+lp_x/16` and `maxmin+lp_x/16` achieve a fairness violation of 0.08 and 0.07 with an overall coverage of 81% and 85% of `grdy_im`'s, respectively – thus strictly dominating over `grdy_maxmin`, `moso` and `myopic`, while beating competitors in terms of fairness. We exclude `ind_lp` as it is not performing too well in terms of fairness and coverage in comparison to `grdy_grp+lp` and `maxmin+lp` for further experiments.

Results for Real World Instances. We turn to the real world instances, see Figure 2 for some results on the networks `Arenas`, `Irvine`, and `email-Eu-core`. Our algorithms `grdy_grp+lp` and `maxmin+lp` achieve the best demographic parity values by far. On the `Arenas` network, for example, we achieve a violation in demographic parity of only 0.008, while getting more than 88% of `grdy_im`'s

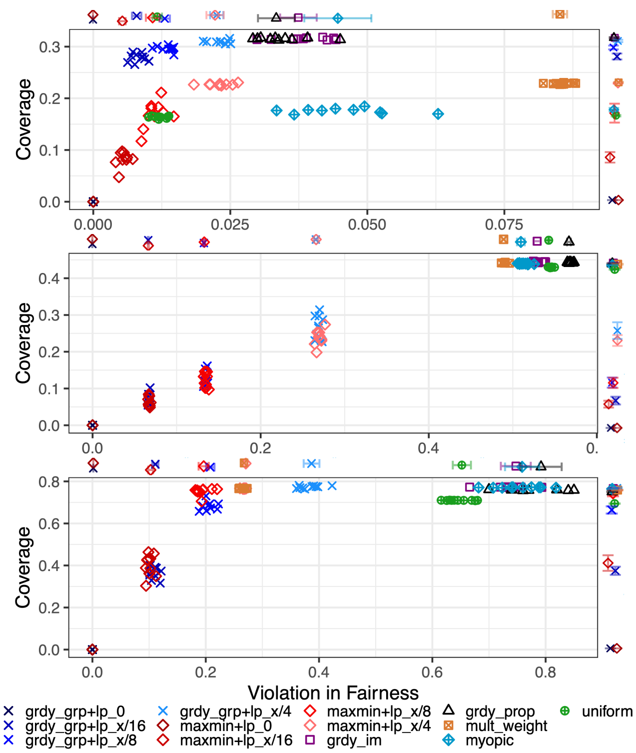

Figure 2: (1) `Arenas` (random-overlap communities, $m = 10, k = 100$), (2) `Irvine` (BFS communities, $m = 10, k = 50$), (3) `email-Eu-core` (real communities, $k = 100$).

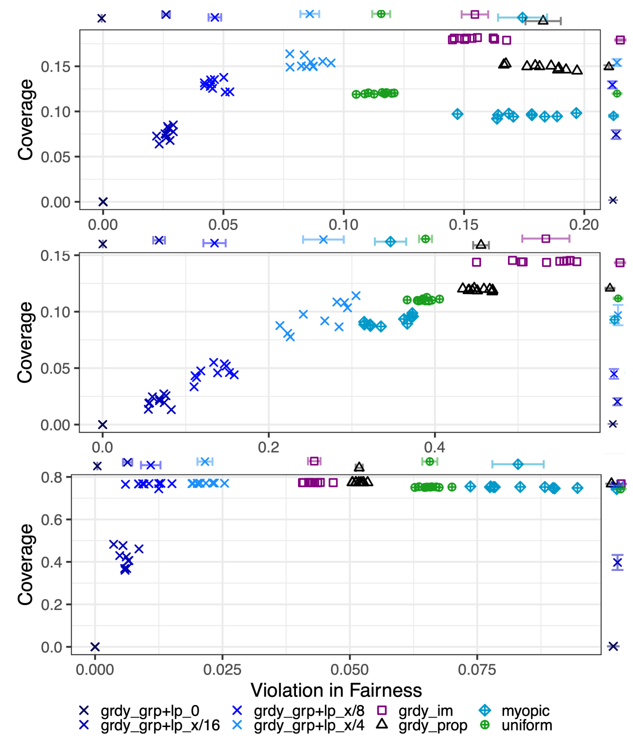

Figure 3: (1) `ca-GrQc` (leidenalg communities, $k = 100$), (2) `ca-HepTh` (random communities, $m = n/10$, $k = 100$), (3) `Facebook` (BFS communities, $m = 2$, $k = 50$).

coverage that in turn suffers an around 5 times higher fairness violation. On the `email-Eu-core` network, our algorithm **`maxmin+lp_x/8`** achieves a fairness violation around 0.2 (a quarter of **`grdy_im`**), while still achieving essentially the same coverage. We note that the simple heuristic **`grdy_prop`** performs even worse in terms of fairness than **`grdy_im`** on the `Irvine` network. We also note that all algorithms but **`grdy_grp+lp`**, **`maxmin+lp`**, and **`mult_weight`** perform comparable to **`uniform`** in terms of both coverage and fairness on `Irvine` and `email-Eu-core`. Lastly, we report on the results for the co-authorship networks `ca-GrQc`, `ca-HepTh`, and the `Facebook` network. Due to running times we further restrict the evaluated algorithms by excluding also **`maxmin+lp`** and **`mult_weight`**. Again **`grdy_grp+lp`** achieves the best fairness values by far. We again see a trade-off between fairness violation and overall coverage, i.e., in some cases no algorithm achieves low fairness violation while maintaining high coverage. Still in some other cases our algorithms achieve exactly that. For `Facebook`, **`grdy_grp+lp_x/16`** obtains 55% of **`grdy_im`**'s coverage with only 7% of its fairness violation. Maybe even better, **`grdy_grp+lp_x/8`** obtains 99% of **`grdy_im`**'s coverage with only 23% of its fairness violation.

Conclusion

We consider the impact of introducing strict demographic parity fairness via constraints in influence maximization through the study of three optimization problems, IM^{dp}, pIM^{dp}, and iIM^{dp}– in an ex-post in case of the former and in an ex-ante fashion in case of the latter two. After showing that this drastically differs from, e.g., the maximin criterion, we studied the price of introducing fairness via constraints in all three problems and observe that it may be unbounded. We then turned to investigating the computational complexity of the three optimization problems and observed that, unless $P = NP$, one cannot approximate IM^{dp} in polynomial time even when the demographic parity fairness constraints are allowed to be violated by a multiplicative or additive term. For pIM^{dp}, we show that the problem is NP-hard, while for iIM^{dp} we even show that it cannot be approximated within a factor better than $1 - 1/e$ unless $P = NP$. We then proposed algorithms for pIM^{dp} and iIM^{dp}. In the case of iIM^{dp} we essentially gave a $1 - 1/e$-approximation algorithm that violates the fairness constraints by at most a $1 - 1/e$-factor as well. For pIM^{dp} we gave two heuristics that allow the user to freely choose the level of tolerated fairness violation. In an extensive experimental study, we then showed that these three algorithms, and particularly the latter two, perform well in practice. That is, for random, synthetic, and real word instances, we obtain the best demographic parity fairness values among all competitors and for certain instances even obtain comparable overall spread. The latter indicates that the *empirical* price of demographic parity fairness may actually be small when using our algorithms in practice.

Acknowledgments

This work was partially supported by the Italian MIUR PRIN 2017 Project "ALGADIMAR" Algorithms, Games, and Digital Markets.

References

Ali, J.; Babaei, M.; Chakraborty, A.; Mirzasoleiman, B.; Gummadi, K. P.; and Singla, A. 2022. On the Fairness of Time-Critical Influence Maximization in Social Networks (Extended Abstract). In *38th IEEE International Conference on Data Engineering, ICDE 2022, Kuala Lumpur, Malaysia, May 9-12, 2022*, 1541–1542. IEEE.

Anwar, M. S.; Saveski, M.; and Roy, D. 2021. Balanced Influence Maximization in the Presence of Homophily. In *WSDM2021*, 175–183. ACM.

Aziz, H.; Brandt, F.; and Stursberg, P. 2013. On popular random assignments. In *SAGT2013*, 183–194. Springer.

Banerjee, A.; Chandrasekhar, A. G.; Duflo, E.; and Jackson, M. O. 2013. The diffusion of microfinance. *Science*, 341(6144): 1236498.

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. Fairness and Machine Learning: Limitations and Opportunities. http://www.fairmlbook.org. Accessed: 2022-11-07.

Becker, R.; Corò, F.; D'Angelo, G.; and Gilbert, H. 2020. Balancing Spreads of Influence in a Social Network. In *AAAI2020*, 3–10.

Becker, R.; D'Angelo, G.; and Ghobadi, S. 2023. On the Cost of Demographic Parity in Influence Maximization. *CoRR*, abs/2302.13110.

Becker, R.; D'Angelo, G.; Ghobadi, S.; and Gilbert, H. 2022. Fairness in Influence Maximization through Randomization. *J. Artif. Intell. Res.*, 73: 1251–1283.

Bogomolnaia, A.; and Moulin, H. 2001. A new solution to the random assignment problem. *Journal of Economic theory*, 100(2): 295–328.

Borgs, C.; Brautbar, M.; Chayes, J. T.; and Lucier, B. 2014. Maximizing Social Influence in Nearly Optimal Time. In *SODA2014*, 946–957.

Brandl, F.; Brandt, F.; and Seedig, H. G. 2016. Consistent probabilistic social choice. *Econometrica*, 84(5): 1839–1880.

Chen, W.; and Teng, S.-H. 2017. Interplay between social influence and network centrality: a comparative study on shapley centrality and single-node-influence centrality. In *WWW2017*, 967–976.

Cohen, E.; Delling, D.; Pajor, T.; and Werneck, R. F. 2014. Sketch-based Influence Maximization and Computation: Scaling up with Guarantees. In *CIKM2014*, 629–638. ACM.

Farnadi, G.; Babaki, B.; and Gendreau, M. 2020. A Unifying Framework for Fairness-Aware Influence Maximization. In *FATES2020 – WWW2020 Companion*, 714–722. ACM / IW3C2.

Fish, B.; Bashardoust, A.; danah boyd; Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. 2019. Gaps in Information Access in Social Networks? In *WWW2019*, 480–490. ACM.

Gershtein, S.; Milo, T.; and Youngmann, B. 2021. Multi-Objective Influence Maximization. In *EDBT*, 145–156.

Katta, A.-K.; and Sethuraman, J. 2006. A solution to the random assignment problem on the full preference domain. *Journal of Economic theory*, 131(1): 231–250.

Kempe, D.; Kleinberg, J. M.; and Tardos, É. 2015. Maximizing the Spread of Influence through a Social Network. *Theory of Computing*, 11: 105–147.

Khajehnejad, M.; Rezaei, A. A.; Babaei, M.; Hoffmann, J.; Jalili, M.; and Weller, A. 2020. Adversarial Graph Embeddings for Fair Influence Maximization over Social Networks. In *IJCAI2020*, 4306–4312.

Machina, M. J. 1989. Dynamic consistency and non-expected utility models of choice under uncertainty. *Journal of Economic Literature*, 27(4): 1622–1668.

Rahmattalabi, A.; Jabbari, S.; Lakkaraju, H.; Vayanos, P.; Izenberg, M.; Brown, R.; Rice, E.; and Tambe, M. 2021. Fair Influence Maximization: a Welfare Optimization Approach. In *AAAI2021*, 11630–11638.

Sadeh, G.; Cohen, E.; and Kaplan, H. 2020. Sample Complexity Bounds for Influence Maximization. In *ITCS2020*, volume 151 of *LIPIcs*, 29:1–29:36.

Stoica, A.; and Chaintreau, A. 2019. Fairness in Social Influence Maximization. In *FATES2019 – WWW2019 Companion*, 569–574. ACM.

Stoica, A.; Han, J. X.; and Chaintreau, A. 2020. Seeding Network Influence in Biased Networks and the Benefits of Diversity. In *WWW2020*, 2089–2098. ACM / IW3C2.

Tang, Y.; Shi, Y.; and Xiao, X. 2015. Influence Maximization in Near-Linear Time: A Martingale Approach. In *SIGMOD2015*, 1539–1554.

Tang, Y.; Xiao, X.; and Shi, Y. 2014. Influence maximization: near-optimal time complexity meets practical efficiency. In *SIGMOD2014*, 75–86.

Traag, V. A.; Waltman, L.; and van Eck, N. J. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1): 5233.

Tsang, A.; Wilder, B.; Rice, E.; Tambe, M.; and Zick, Y. 2019. Group-Fairness in Influence Maximization. In *IJCAI2019*, 5997–6005.

Wang, X.; Varol, O.; and Eliassi-Rad, T. 2021. Information Access Equality on Network Generative Models. *CoRR*, abs/2107.02263. Available at SSRN.

Wilder, B.; Onasch-Vera, L.; Hudson, J.; Luna, J.; Wilson, N.; Petering, R.; Woo, D.; Tambe, M.; and Rice, E. 2018a. End-to-End Influence Maximization in the Field. In *AAMAS2018*, 1414–1422.

Wilder, B.; Ou, H.; de la Haye, K.; and Tambe, M. 2018b. Optimizing Network Structure for Preventative Health. In *AAMAS2018*, 841–849.

Wu, Q.; Li, Z.; Wang, H.; Chen, W.; and Wang, H. 2019. Factorization Bandits for Online Influence Maximization. In *KDD2019*, 636–646. ACM.

Yadav, A.; Wilder, B.; Rice, E.; Petering, R.; Craddock, J.; Yoshioka-Maxwell, A.; Hemler, M.; Onasch-Vera, L.; Tambe, M.; and Woo, D. 2018. Bridging the Gap Between Theory and Practice in Influence Maximization: Raising Awareness about HIV among Homeless Youth. In *IJCAI2018*, 5399–5403.