IM-Balanced: Influence Maximization Under Balance Constraints

Shay Gershtein, Tova Milo, Brit Youngmann and Gal Zeevi
Tel Aviv University
{shayg1,milo,brity,galzeevi}@post.tau.ac.il

ABSTRACT

Influence Maximization (IM) is the problem of finding a set of influential users in a social network, so that their aggregated influence is maximized. IM has natural applications in viral marketing and has been the focus of extensive recent research. One critical problem, however, is that while existing IM algorithms serve the goal of reaching a large audience, they may obliviously focus on certain well-connected populations, at the expense of key demographics, creating an undesirable imbalance, an illustration of a broad phenomenon referred to as *algorithmic discrimination.* Indeed, we demonstrate an inherent trade-off between two objectives: (1) maximizing the overall influence and (2) maximizing influence over a predefined “protected" demographic, with the optimal balance between the two being open to different interpretations. To this end, we present IM-Balanced, a system enabling end users to declaratively specify the desired trade-off between these objectives w.r.t. an emphasized population. IM-Balanced provides theoretical guarantees for the proximity to the optimal solution in terms of both objectives and ensures an efficient, scalable computation via careful adaptation of existing state-of-the-art IM algorithms. Our demonstration illustrates the effectiveness of our approach through real-life viral marketing scenarios in an academic social network.

1 INTRODUCTION

Social networks attracting millions of people, such as Facebook, LinkedIn, and Sina Weibo, have emerged recently as a prominent marketing medium. *Influence Maximization* (IM) is the problem of finding a set of influential users (termed a *seed set*) in the network, so that their aggregated influence is maximized [7]. IM has a natural application in viral marketing, where companies promote their brands through the word-of-mouth propagation. This has motivated extensive research [2, 9], emphasizing the development of scalable IM algorithms [6, 11].

The majority of IM works focus on maximizing the overall influence, given a seed set size requirement. While this serves the goal of reaching a large audience, IM algorithms may obliviously focus on certain well-connected populations, at the expense of key demographics. This may create an undesirable imbalance, a conspicuous illustration of a broad phenomenon referred to as *algorithmic discrimination.* In this work, we assume the existence of a boolean function(s) over user profile attributes, which identifies a *protected user group(s).* This function can model juridical definitions of protected demographics, or be any arbitrary boolean query over multiple attributes.

As an example, consider a high-tech company interested in recruiting researchers, opting for a social media campaign to inform as many candidates as possible, through opinion leaders in the field. Employing an IM algorithm may produce a campaign overlooking a protected group of users, e.g., characterized by gender, age and/or country. This latent, non-meritocratic discrimination harms both potential candidates and the company, impeding the promotion of a balanced environment and creating a vulnerability to discrimination lawsuits. With the pervasiveness of such automated processes, these concerns have substantial economic and moral repercussions.

A related line of research is concerned with *targeted IM* algorithms [10], which aim to find a seed set maximizing the influence over users relevant to a given topic/context [9]. While these works provide theoretical guarantees for this objective, they do not ensure that the overall influence is sufficiently large, compared to a non-targeted IM algorithm. Continuing with our example, while promoting the exposure to protected users is important, general large exposure is still essential to identify the best candidates, and is not guaranteed by existing targeted IM algorithms. Indeed, as we show in this work, given a seed set size requirement, there exists an inherent trade-off between the two objectives: (1) maximize the overall influence and (2) maximize influence over protected users; hindering a simultaneous optimization of both (see Section 2.2). This trade-off produces a spectrum of different possible combinations of emphases on each objective, with the particular choice being application dependent, corresponding to different interpretations of balance, fairness and diversity in the literature [3, 12, 13].

To the best of our knowledge, IM-Balanced is the first system enabling users to declaratively specify the desired trade-off between the two objectives. IM-Balanced allows users to specify the protected population and the notion of balance between the objectives that they wish to achieve. Our algorithm ensures an efficient, scalable computation, while providing theoretical guarantees for the proximity to the optimal solution in terms of both objectives. For simplicity of presentation, we assume in the next sections a single boolean function defining one protected group, but our definitions, results and system apply to multiple functions corresponding to possibly overlapping subsets of the populations.

The two key challenges that our system tackles are as follows.

Balanced IM. What is the correct trade-off between the objectives? As the definition is arguably subjective and context dependent, it requires a flexible, tunable system that enables to explicitly manage this trade-off. IM-Balanced allows the user to prioritize the objectives and transform one into a parametrized constraint. For example, “Maximize the overall influence, while ensuring the influence over protected users is above a given threshold", or, alternatively, “Maximize the influence over protected users, while ensuring the overall influenced is above a given threshold". That is, IM-Balanced enables a *tunable definition of balance* (to be formally defined in Section 2), where the user can declaratively specify: (1) The required size of the seed set; (2) The protected group of users; (3)

the balancing criteria, which translates to customizable objective and constraint functions, along with a size threshold parameter.

Efficiency and Scalability. State-of-the-art IM algorithms are the product of decades of research focused on scalability, capable of managing billion-node networks. A key challenge is, thus, to provide an algorithm on par with existing IM algorithms, in terms of performance. Given a user specification, IM-Balanced generates an *algorithm instance* suiting the user-defined notion of balance. The generated instances employ existing IM algorithms as a white box, with minor adaptations, thus supporting extensibility and facilitating performance comparable with top performing IM algorithms. IM-Balanced is also assured to satisfy the constraint while providing theoretical guarantees for the objective.

Demonstration overview. We demonstrate the operation of IM-Balanced through the reenactment of a real-life viral marketing scenario, where the system is used to identify influential individuals in the research community, for the purpose of recruiting researchers for a high-tech company. For our illustration, influence is captured in terms of citations and collaborations, inferred from a real-life academic social network [1], and protected subpopulations are illustratively defined in terms of various attributes such as gender and nationality. We first present to the audience several examples of subpopulations that are indeed neglected by standard IM algorithms, alongside results obtained by our system, demonstrating the advantages of our approach. Beyond the intended recruitment, the obtained results will also allow us to highlight trends and patterns in the research community, and, in particular, identifying isolated demographics and unintended imbalances, along with suggesting impactful focal points for a corrective campaign. The audience will actively participate in the demonstration by selecting their desired protected groups and composing various balance definitions, then reviewing the results in contrast to those obtained by other baseline approaches. See more details in Section 3.

2 TECHNICAL BACKGROUND

We start by providing a brief overview of the standard Influence Maximization (IM) problem, then present our framework for Balanced-IM. For space constraint, proofs and additional examples are deferred to our technical report [5].

2.1 Influence Maximization

We first recall the definition of IM, then briefly overview its complexity and its top performing algorithm schema. We model a social network as a directed weighted graph $G = (V, E, W)$, where V is the set of nodes and each edge $(u, v) \in E$ is associated with a weight $W(u, v) \in [0, 1]$, which models the probability that node u will influence its neighbor v. Given a function $I(\cdot)$ dictating how influence is propagated in the network and a number k, IM is the problem of finding a seed set $O = argmax_{T \subseteq V, |T|=k} I(T)$, where $I(T)$ denotes the expected number of nodes influenced by T. The function $I(\cdot)$ is defined by an *influence propagation model.* The majority of existing IM algorithms apply for the *Independent Cascade* (IC) and the *Linear Threshold* (LT) models [2, 6], both proposed in [7]. Our results hold under both models, but, for simplicity of presentation, we focus on the IC model.

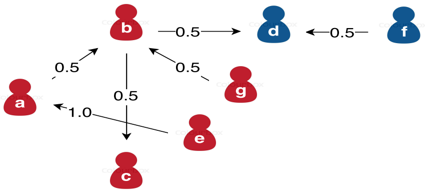


Figure 1: Example colored social network.

We refer to influenced nodes as *covered.* Initially only seed nodes are covered. In the IC model the propagation is carried out in discrete steps, s.t. each node covered in the preceding step attempts to influence its uncovered neighbors, with an independent probability indicated by the weight of the edge connecting them.

Selecting the optimal seed set is NP-hard, with inapproximability beyond a factor of $(1 - \frac{1}{e})$ [7]. Recent IM algorithms, based on the Reverse Influence Sampling (RIS) approach [2], achieve optimal accuracy in near optimal time [11]. The RIS framework samples nodes independently and uniformly, then for each sampled node, constructs a Reverse Reachability (RR) set consisting of its sampled sources of influence. Next, the problem is reduced to an instance of the *Maximum Coverage* problem, where k nodes are selected with the goal of maximizing the number of covered RR sets.

2.2 Balanced IM

As mentioned, our framework supports multiple (possibly overlapping) protected groups, however, for simplicity, we assume here a single such group. In our setting, nodes (users) are colored s.t. a subset of users (blue nodes) belong to a protected group, with all other users (red nodes) referred to as non-protected. For example, Figure 1, depicts a sample colored network. Recall that $I(S)$ denotes the expected number of covered users by the seed set S, and let $I_r(S)$, $I_b(S)$ denote the expected number of red/blue users covered by S, resp. Additionally, let O denote the optimal k-size solution in terms of cover size. For example, in Figure 1, for $k = 2$, $O = \{e, g\}$, $I(O) = 4\frac{1}{2}$, $I_r(O) = 4\frac{1}{8}$ and $I_b(O) = \frac{3}{8}$. One can see that O covers almost exclusively red users.

To obtain a more balanced solution, one may request that a larger number of blue nodes should be covered. However, this requirement alone, if not properly constrained, may lead to a drastic decrease in the overall cover size, rendering the solution undesirable. To illustrate, consider again Figure 1 with $k = 2$. As mentioned, the size of the optimal solution in terms of cover size is $I(O) = 4\frac{1}{2}$ and $I_b(O) = \frac{3}{8}$. Nonetheless, the optimal solution in terms of covered blue nodes is $B = \{d, f\}$, where $I(B) = I_b(B) = 2$ and $I_r(B) = 0$, which covers a greater number of blue nodes at the cost of significantly reducing the overall cover.

This simple example exposes the inherent trade-off between these two objectives, implying that instead of naively maximizing both simultaneously, one should prioritize the objectives and transform the secondary objective into a constraint strong enough to ensure balance, but weak enough to provide a necessary degree of freedom in optimizing the main objective. Towards this end, IM-Balanced enables the user to specify: (1) The number k of seed nodes; (2) The protected group of users; (3) The objective and the constraint functions, and (4) The threshold parameter $t \in [0, 1]$, that restricts the extent to which the solution is allowed to deviate from the optimum for the constraint.

To illustrate, consider the following example definition of Balanced IM, referred to as the *protected-oriented* definition: Given the parameters k and t, find a seed set B^* that maximizes the number of covered blue nodes, subject to a constraint on the overall cover size being above the specified fraction of its optimal (possibly unbalanced) maximal value. Namely,

$$\boxed{B^* = argmax_{|T|=k, I(T)\geq t\cdot(1-\frac{1}{e})\cdot I(O)} I_b(T)}$$

Recall that O is the optimal solution in terms of cover size. Note that in the above formula the expected cover size of a given set is compared to $(1-\frac{1}{e})\cdot I(O)$, rather than to $I(O)$, since even for standard IM, unless $P = NP$, no polynomial algorithm can guarantee a cover size greater than $(1-\frac{1}{e})\cdot I(O)$ [7].

Similarly, one can choose to maximize the overall cover size, subject to the constraint that enough blue nodes are covered. We refer to this definition as *size-oriented.* Namely, find a set O^* satisfying:

$$\boxed{O^* = argmax_{|T|=k, I_b(T)\geq t\cdot(1-\frac{1}{e})\cdot I_b(B)} I(T)}$$

where B is the optimal k-size solution in terms of covered blue nodes. Here again, we compare $I_b(T)$ to $(1-\frac{1}{e})\cdot I_b(B)$ rather than to $I_b(B)$, as we can prove that the same complexity bound mentioned above holds for this variation as well. The user can similarly choose other balance definitions that, e.g. constrain the number of covered red nodes, enforce a minimal ratio of blue to red covered nodes, or add constraints on the selected seed nodes.

2.3 Computing the Balanced IM solution

Given a user specification, IM-Balanced generates an *algorithm instance*, suited for that notion of balance. As mentioned, the instance employs, as a white-box, an existing RIS-based IM algorithm (e.g., [6, 11]). We start by shortly describing our generic modification of a given IM algorithm, followed by our full solution scheme.

Protected-aware IM. Given an (RIS based) algorithm $\mathcal{A}$, we define $\mathcal{A}_b$ to be its protected-aware version, i.e., while $\mathcal{A}$ maximizes the overall cover size, $\mathcal{A}_b$ maximizes the number of covered blue nodes exclusively. Any RIS-based algorithm can be adapted to its protected-aware counterpart via a single modification: the RR sets are sampled from blue nodes only. We can prove that $\mathcal{A}_b$ outputs a solution covering at least $(1-\frac{1}{e})\cdot I_b(B)$ blue nodes, where B is the optimal k-size solution maximizing $I_b(\cdot)$ [5]. Analogously, we define $\mathcal{A}_r$ to be a variant of an IM algorithm $\mathcal{A}$, which maximizes the influence over the red nodes exclusively.

Generating a definition-dedicated algorithm instance. To ease the presentation, we first illustrate the algorithm template generated for the *protected-oriented* balance definition, then briefly explain how this generalizes to support customizable alternative definitions.

The algorithm instance IM-Balanced generates for this definition is depicted in Algorithm 1. It runs independently two procedures: one ensures that the solution satisfies the constraint[1] (line 3.1), and the second maximizes the objective (line 3.2). It then returns the union S of the selected seeds. If S contains less than k seeds, it runs $\mathcal{A}_b$ on the residual problem to complete the seed set (lines 5-7).

Recall that O^* denotes the k-size optimal seed set for the protected-oriented definition. We can prove that Algorithm 1 guarantees a $(1-t)\cdot(1-\frac{1}{e})$-approximation to the protected-oriented definition. That is: $I_b(S)\geq(1-t)\cdot(1-\frac{1}{e})\cdot I_b(O^*)$ and $I(S)\geq t\cdot(1-\frac{1}{e})\cdot I(O)$. Note that the time complexity of the algorithm depends on that of $\mathcal{A}$ (we run $\mathcal{A}$ twice), which is nearly optimal [6, 11].

Algorithm 1 Algorithm instance for the protected-oriented balance definition.

```
1: Input: The parameters $k$ and $t$ and an algorithm $\mathcal{A}$.
2: Output: A $k$-size solution $S$.
3: We run independently the following two procedures:
       (1) $S_1 \leftarrow$ Run algorithm $\mathcal{A}$ with $k' = \lceil t\cdot k\rceil$.
       (2) $S_2 \leftarrow$ Run algorithm $\mathcal{A}_b$ with $k' = \lfloor(1-t)\cdot k\rfloor$.
4: $S \leftarrow S_1 \cup S_2$
5: if $|S| < k$ then
6:     Run $\mathcal{A}_b$ again until $k$ seeds are gathered.
7: end if
8: return $S$
```

Finally, we conclude with a brief explanation of the algorithm instances generated for other balance definitions. Conceptually, given a balance definition, all that needs to be adjusted is the number of seeds required for each of the algorithms $\mathcal{A}$, $\mathcal{A}_b$ and $\mathcal{A}_r$. For example, to comply with the *size-oriented* definition, we set algorithms $\mathcal{A}$ and $\mathcal{A}_b$ to return $\lceil(1-t)\cdot k\rceil$ and $\lfloor t\cdot k\rfloor$ seeds, resp. As another example, one may ask to maximize the number of covered blue nodes, subject to a constraint on the number of covered red nodes. To support this definition, we run $\mathcal{A}_b$ and $\mathcal{A}_r$ to return $\lceil(1-t)\cdot k\rceil$ and $\lfloor t\cdot k\rfloor$ seeds resp. For more details see [5].

3 SYSTEM AND DEMONSTRATION

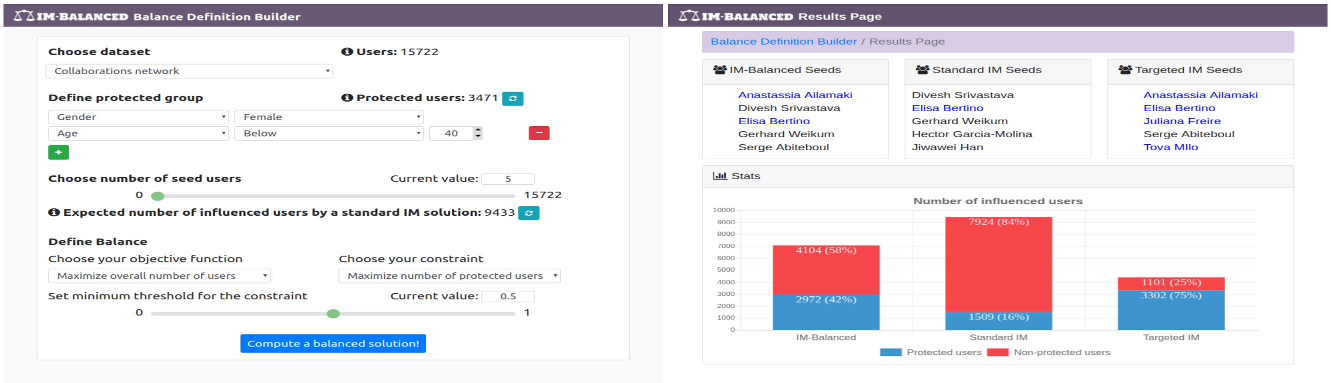
Implementation. IM-Balanced is implemented in Python 2.7. The IM algorithm our prototype employs is IMM [11]. The user specifies her balance definition via the UI (to be described next), implemented in HTML5/CSS3, then the system runs our generic algorithm to generate, and efficiently run, the suitable algorithm instance. The results are displayed on a results page (also described next), along with charts depicting various statistics. The user may examine and compare results, and correspondingly refine her inquiry.

Demonstration. As mentioned, we demonstrate the capabilities of IM-Balanced through the reenactment of a real-life viral marketing scenario, where the system is used for the advertisement of open research positions in a high-tech company.

For our illustration, we have constructed an influence graph based on a social network of researchers extracted from [1], focusing on the Database and Data Mining community. The profile of a researcher, used to define the protected subpopulations, includes details about her gender, country, age, etc. The mutual influence relationships capture information on the researchers' past collaborations and citations (with weights reflecting the portion of collaborations/citations between two authors).

In our first demonstration scenario, we assume (for the sake of illustration) that the company wants to ensure that a large number of researches are informed about the opening, while guaranteeing that enough young female researchers are informed as well. Using the system's UI, as depicted in Figure 2a (from top to bottom), CIKM participants will choose: (1) which part of the network to consider (only the collaboration edges, only citations, or both); (2) the properties of the protected group - here, females under 40 (the cardinality of this group is displayed in real time); (3) the required

[1] The rounding operation (ceiling) ensures that the output satisfies the constraint and the number of seeds algorithm $\mathcal{A}$ returns is an integer.



IM-BALANCED Balance Definition Builder

Choose dataset

Collaborations network

Users: 15722

Define protected group

Protected users: 3471

Gender | Female

Age | Below | 40

Choose number of seed users

Current value: 5

0 — 15722

Expected number of influenced users by a standard IM solution: 9433

Define Balance

Choose your objective function: Maximize overall number of users

Choose your constraint: Maximize number of protected users

Set minimum threshold for the constraint

Current value: 0.5

0 — 1

Compute a balanced solution!

(a) Balance definition builder.

IM-BALANCED Results Page

Balance Definition Builder / Results Page

IM-Balanced Seeds	Standard IM Seeds	Targeted IM Seeds
Anastassia Ailamaki	Divesh Srivastava	Anastassia Ailamaki
Divesh Srivastava	Elisa Bertino	Elisa Bertino
Elisa Bertino	Gerhard Weikum	Juliana Freire
Gerhard Weikum	Hector Garcia-Molina	Serge Abiteboul
Serge Abiteboul	Jiwawei Han	Tova MIlo

Stats

(b) Results page.

Figure 2: IM-Balanced UI.

seed set size k (the expected influence of a standard IM algorithm for this k is also displayed); (4) a balance definition: the objective - here, maximize the overall influence, and the constraint - the size of the protected cover is at least 50% of the size of the optimal cover of protected users ($t = 0.5$). Pressing the "Compute a balanced solution" button, the system runs the instantiated algorithm. The results page, as portrayed in Figure 2b, is then presented.

As one of the key objectives of this demonstration is to enable the audience to gauge the effectiveness of IM-Balanced, the results are displayed alongside those of previous IM algorithms that either focus solely on maximizing the overall influence [11] or alternatively focus on the protected group alone (targeted-IM) [9]. To illustrate, Figure 2b depicts such a comparison, showing that the solution returned by IM-Balanced influences almost the same number of users as the standard IM algorithm that focuses only on overall size ($7,076$ vs. $9,433$), while influencing almost twice as many protected users. We can see in the figure that IM-Balanced also comes very close to the targeted-IM solution w.r.t. the influenced protected users ($2,972$ vs. $3,302$), while influencing overall significantly more users, and thus is clearly more advantageous. Additionally, the selected seeds (ordered alphabetically) of each algorithm are presented to the participants, allowing for the discovery, for instance, of which researches have significant influence on young women while at the same time influencing the community at large (compared, e.g., to those influencing mostly a large male population).

The audience will then be invited to formulate other balance notions by tuning the protected group definition (e.g. researchers form undeveloped countries) and the balance criteria (e.g. switch between the objective and the constraint), and will examine how these affect the results.

Finally, to have the audience further experience the system, we will present to the participants other viral marketing tasks on this network, such as calls for nominations for awards/grants applications. In each scenario, we will examine various protected populations which are neglected by standard IM algorithms, consider several balancing criteria, and correspondingly examine how these affect the selected seeds and the size/type of the influenced populations. Last, to demonstrate the robustness of IM-Balanced, interested participants will be further allowed to examine the system's operation on additional real-life social networks such as Pokec, a popular social network in Slovakia and data extracted from Twitter [8], whose graph datasets were also ported into the system.

4 RELATED WORK

The study of *algorithmic discrimination*, *fairness* and *diversity* has been gaining popularity in recent years. Work on *fairness*, with the aim of remedying algorithmic bias against groups or individuals on unreasonable grounds, focused largely on predictive tasks [12] and ranking [3, 13]. *Diversity*, i.e., ensuring that different kinds of objects are represented in the output of an algorithm as opposed to similar high-scoring results, has been studied in the context of search engines and recommender systems [4]. Each of these concepts is naturally subject to various context-dependent interpretations and plays a different role in studying algorithmic imbalances, hence the parametrization of our framework to accommodate a variety of applications. Our definition of *protected group* generalizes the definitions used in previous work [4, 13], by capturing both binary and non-binary attributes of possibly overlapping subsets of users.

As mentioned, IM has been studied extensively, with emphasis on scalable performance [6, 7]. IM-Balanced can employ any top performing IM algorithm (e.g., [11]) in a white-box manner, match its state-of-the-art performance and take advantage of all its optimizations (e.g., parallelized computation), while also retaining theoretical accuracy guarantees.

REFERENCES

[1] AMiner 2018. AMiner. (2018). https://aminer.org/data.

[2] Chayes J. Borgs C., Brautbar M. and Lucier B. 2014. Maximizing Social Influence in Nearly Optimal Time. In *SODA*.

[3] Straszak D. Celis L E. and Vishnoi N K. 2017. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840* (2017).

[4] Pitoura E. Drosou M., Jagadish HV and Stoyanovich J. 2017. Diversity in big data: A review. *Big data* (2017).

[5] full version 2018. Balanced IM: Technical Report. (2018). https://goo.gl/2GZxZk.

[6] B. Glenn X. Xiaokui Huang K., W. Sibo and Lakshmanan L. V. S. 2017. Revisiting the Stop-and-stare Algorithms for Influence Maximization. *PVLDB* (2017).

[7] Kleinberg J. Kempe D. and Tardos E. 2003. Maximizing the Spread of Influence Through a Social Network. In *KDD*. ACM.

[8] Jure Leskovec and Andrej Krevl. 2014. SNAP Datasets: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data. (June 2014).

[9] Yuchen Li, Dongxiang Zhang, and Kian-Lee Tan. 2015. Real-time Targeted Influence Maximization for Online Advertisements. *PVLDB* (2015).

[10] Chonggang Song, Wynne Hsu, and Mong Li Lee. 2016. Targeted Influence Maximization in Social Networks. In *CIKM*. ACM.

[11] Shi Y. Tang Y. and Xiao X. 2015. Influence Maximization in Near-Linear Time: A Martingale Approach. In *SIGMOD*.

[12] Gomez R. M. Zafar M. B., Valera I. and Gummadi K. P. 2017. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259* (2017).

[13] Castillo C. Hajian S. Megahed M. Zehlike M, Bonchi F. and Baeza-Yates R. 2017. Fa* ir: A fair top-k ranking algorithm. In *CIKM*. ACM.