Equity of Reproduction: A reproduction of Amortizing Individual Fairness in Rankings

Emil Dudev
emil.dudev@student.uva.nl
12767492

Sietze Kuilman
Skkuilman@gmail.com
11930098

Nils Lehmann
nils.lehmann@student.uva.nl
12868175

Thomas van Zwol
T.j.vanzwol@gmail.com
10555714

Marco Heuvelman
makoheuvelman@hotmail.com
-

ABSTRACT

The world of computer science is catching up when it comes to biases. One such problem in bias is an unfair treatment of certain groups. The problem of fairness, however, often involves contradictory ideas. One paper introduces a version of fairness in the world of ranking systems, namely: the notion of proportional attention. It assumes that attention is a resource which should be divided proportionally to the relevance one contributes. A reproduction of this paper was attempted. The algorithm proved to be reproducible. Yet due to a lack of information about the datasets, a lack of statistical tests, and unclear design choices, we were unable to accurately reproduce the results. However, the results do show that the algorithm treats people more fairly. Whether this notion of fairness is actually fair of course is still up for debate.

KEYWORDS

Algorithmic Fairness, Fair Ranking, Position Bias, Exposure

1 INTRODUCTION

> *"We can affirm the unavoidable use of technical devices, and also deny them the right to dominate us, and so to warp, confuse, and lay waste our nature."* - Martin Heidegger, Discourse on Thinking, p.54 1969 [1].

In today's world we use machines to order the veritable jungle of information. In yesteryear we relied on scores of librarians who ordered the world of information [2]. Yet with the staggering amount of unstructured data today, this is a task far beyond manual capabilities. The systems we use to rank and the data that goes along with them, however, are not without their problems.

At the turn of the century questions surrounding the political aspects of search engines arose [3]. By it's very nature such ranking systems create systemic problems. Making choices means introducing a certain bias, we cannot do without. The scale and the resolution on which algorithms can function, however, may make these biases problematic. Liu et al. [4] mention the harm that automated decision procedures may introduce. Without the human factor we are delivered over to a system without any form of ethics or leeway. Without a notion of fairness we may introduce, unconscious or otherwise, systemic biases towards certain groups. In other words there is a trade-off, we need to introduce some measure of bias to create meaningful decisions but by minimizing the bias on sensitive features we can create a more "fair" ranker [5]. The world of computer science is catching up and we can see a research paradigm arise which is trying to incorporate notions of fairness within automated decision making [6, 7].

Of course the important question here is: what is "fair"? The definitions are often contradictory and political. The arbitrary distribution Nozick [8] favours can be viewed as fair but to others this sounds ludicrous [9] . The same goes for the Rawlsian notions of improving the opportunities for the least advantaged party [10]. The introduction of fairness within the world of ranking systems should be considered political as Introna and Nussbaum "happily" point out [3]. The preference over one piece of information over another is a political act according to them.

In this paper we attempt to reproduce a paper pertaining to one such notion of fairness within the world of ranking systems namely: The notion of proportional attention [11]. It is based on Dworks et al. [12] and assumes that similar individuals should be treated similarly. The resource to be fairly distributed is assumed to be attention. Their notion of proportionality attempts to create some form of equity of attention.

We will scrutinize the specifics of the paper itself. In the method section we will look at the original paper and distill the essence. In the experimental setup we show the specific implications and how we differ from the original. The rest of the paper discusses the results and what implication the paper has on the field at large.

2 METHOD

The aim of Biega et al. [11] is to ensure individual fairness across multiple rankings. They state that individual fairness is impossible in the context of only a single ranking, but claim their approach of amortizing the attention across rankings results in more fairness. The concept is the following: attention is a resource that has to be divided fairly among subjects. They state that a fair way to divide this, is by looking at how worthy a subject is of attention. Relevance is considered to be the best proxy of worthiness. The attention a subject in a ranking receives depends on the position, which is determined by the relevance for the current query. [1] However, the amount of attention per position is not proportional to the relevance of the subjects in the ranking. This results in the top positions receiving more attention than they deserve and other positions receiving less than they deserve. If the difference between the received attention and the amount of relevance contributed to all the rankings up to this ranking becomes too large, the subject

[1] Beiga et al. assume attention depends on position but they make no argument as to why this is the case. Of course there is the golden triangle [13], we could make an argument against this but position bias is generally accepted

will be placed higher (when not having received enough attention) or lower (when having received too much) in the ranking, in order to reduce this imbalance between contributed relevance and received attention.

2.1 Notation

We follow the notation as it has been defined by the authors in [11], being:

- $u_1, ..., u_n$ is a set of subjects ranked in a system,
- $\rho^1, ..., \rho^m$ is a sequence of rankings,
- r_i^j is the [0..1]-normalized relevance score of subject u_i in ranking ρ^j,
- a_i^j is the[0..1]-normalized attention value received by subject u_i in ranking ρ^j,
- A denotes the distribution of accumulated attention across subjects, that is, $A_i = \sum_{j=1}^m a_i^j$ for subject u_i,
- R denotes the distribution of accumulated relevance across subjects, that is, $R_i = \sum_{j=1}^m r_i^j$ for subject u_i,
- w_k is the attention value that is received by a subject at position k in ranking ρ^j.

2.2 Fairness metric

(Un)fairness is measured by calculating the L1-norm: the distributions of A and R should be the same and the L1-norm provides a measure of difference between the two distributions: how much they differ is giving us a unfairness metric. They mention that they could also have used KL-divergence but provide no reasons for their choice.

The formula for the L1-norm is:

$$unfairness(\rho^1, ...\rho^m) = \sum_{i=1}^{n} |A_i - R_i| = \sum_{i=1}^{n} \left| \sum_{j=1}^{m} a_i^j - \sum_{j=1}^{m} r_i^j \right| \quad (1)$$

The formula for the KL-divergence is:

$$D_{KL}(P \parallel Q) = \sum_{x \in X} P(x) \cdot log \frac{P(x)}{Q(x)} \quad (2)$$

Although the authors say that KL-divergence is an alternative unfairness metric, they do not specify how it could be used. The metric is by definition asymmetric (seen by the above formula), and therefore when implementing the algorithm, one has to think how to use the accumulated attention A and the accumulated relevance R.

In machine learning, the KL-divergence metric is interpreted as the information gain if Q is used instead of P. This might hint at the possibility of using the attention values A as Q and the relevance R as P, as we are interested in ranking by (accumulated) attention, and not by relevance. However, the Kullback–Leibler divergence is defined only if for all x, $Q(x) = 0$ implies $P(x) = 0$. This would mean that if for any subject our accumulated attention were to be zero (most common for low-ranking subjects), then the accumulated relevance would have to be zero, as well. This is highly dependant on the data and is not true in the general case.

Moreover, if we were to take the opposite assignment in the formula, namely treat A as P and R as Q, being subject to the definition constraint, we would have to guarantee that for all subjects, if the accumulated relevance were to be zero, then the accumulated attention would also have to be zero. This case has little meaning due to the above-mentioned interpretation of the metric as the information gain. Additionally, it is also invalid in the general case, albeit less likely.

2.3 Ranking quality metric

When a ranking is changed to make it more fair, the ranking may become of lower quality. Thus a measure of ranking quality is also needed, in order to capture the trade-off between quality and fairness. In [11] the NDCG quality metric is used. In NDCG the ground-truth ranking is used to normalize the DCG score of the new ranking. Biega et al. consider the original ranking based on relevance as the ground-truth and thus use the DCG score for the original ranking as the normalization factor[2]. We use the same quality metric in order to ensure the best chance of replicating their results. Other options to evaluate the ranking quality might work just as well, they give comparison between the original ranking and the ranking that is permuted to increase its fairness.

2.4 Optimizing the trade-off between quality and fairness

For this algorithm, a choice has to be made: either minimize unfairness under a constraint of a minimum quality or maximize quality under a constraint of minimum fairness. Biega et al. [11] focused on minimizing unfairness, due to ranking quality measures being more interpretable.

This optimization can be defined as the authors note as a integer linear program (ILP). For the reasoning and explanations we refer the reader to [11], but we will state the specifications of the ILP here.

There are n^2 binary decision variables $X_{i,j}$ for the ILP: a variable is set to 1 when subject u_i is assigned to position j and set to 0 otherwise.

The constants of the ILP for ranking ρ^l are: r_i^l, R_i^{l-1} and A_i^{l-i} for each u_i in ρ^l, w_k for each k, and the IDCG@k value for ρ^l. For all u_i, the accumulated attention and relevance are initialised at zero: $A_i^0 = 0$ and $R_i^0 = 0$.

The ILP is then defined as:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{n} \sum_{j=1}^{n} |A_i^{l-1} + w_j - (R_i^{l-1} + r_i^l)| \cdot X_{i,j} \\ \text{subject to} \quad & \sum_{j=1}^{k} \sum_{i=1}^{n} \frac{2^{r_i^l} - 1}{log_2(j+1)} X_{i,j} \geq \theta \cdot IDCG@k \\ & X_{i,j} \in \{0, 1\}, \forall_{i,j} \\ & \sum_{i} X_{i,j} = 1, \forall_j \\ & \sum_{j} X_{i,j} = 1, \forall_i \end{aligned} \quad (3)$$

[2] We assume the reader is familiar with DCG and NDCG, due to its wide use in the field of IR. In the case it is unknown, Biega et al. [11] provide a concise explanantion.

2.5 Pre-filtering

To reduce the load on the linear solver and speed up the algorithm, pre-filtering is used: the number of subjects that will be considered for reranking in a subset of the total set of subjects in the ranking. This reduces the number of variables that the linear solver has to consider.

Let the size of the subset D be t and the rank at which the quality constraint is calculated k. Then the pre-filtered candidates are the k subjects with the highest relevance score (to make sure the quality constraint can always be satisfied) combined with the $t - k$ subjects with the smallest values for $A_i - (R_i + r_i)$. Because the values for w_j are always positive, the subjects with the smallest values for $A_i - (R_i + r_i)$ are most worthy for being promoted to higher positions. This reduces the number of binary decision variables from n^2 to t^2.

For a more detailed description of the reasons for doing pre-filtering, we refer the reader to the original paper.

One thing to note though, is that Biega et al. do not describe what to do with the subjects that are left out of the pre-filtering. Since we are not really presenting a ranking to users, we didn't try to combine all the subjects into one ranking. What only matters for this research, is which subjects receive attention: these subjects accumulate attention, all the other subjects only accumulate relevance. How much relevance each subject accumulates is not dependant on their rank, only on their relevance score. However, in the library we provide, there is an extra API method, which one can use to generate a complete reranking (not just the unfairness score). We do not use this API, as it is not needed for reproducing the experiments of the original paper.

2.6 Implementation

Our implementation is done in Python and uses an interfacing library[3] to send the linear programming problems to a dedicated linear solver[4]. We used Gurobi through PuLP. The original paper simply refered to using Gurobi[5].

3 EXPERIMENTAL SETUP

For a description of how Biega et al. [11] conducted their experiment, we refer the reader to the original paper. We will discuss interpretations of their approach, and differences between our setup and theirs.

An important thing to note is that the relevance scores for each dataset should be normalized so that they sum to one: to use the L1-norm, or another measure of fairness depending on the difference between two distributions, the relevance for a dataset should be a probability distribution. This is briefly mentioned in the original paper, while being an important cornerstone of their framework.

3.1 Data

3.1.1 Synthetic datasets.

Uniform: The uniform dataset we constructed consists of 100 subjects. The relevance scores for every subject are equal. They are also normalised so that the total relevance sums to one.

City	# Subjects	Date Compiled
Boston	3924	6/10/2017
Geneva	2209	17/11/2017
Hong Kong	4513	7/8/2016

Table 1: Number of subjects within the AirBnB datasets

Linear: The linear dataset we constructed consists of 100 subjects. The relevance scores decrease linearly from the first subject to the last subject. The relevance of all the subjects is normalised so that the total relevance sums to one. Due to the normalisation of the dataset, the slope of the linear formula describing the dataset is fixed, so this dataset matches the one used by the original authors exactly.

Exponential: The exponential dataset we constructed consists of 100 subjects, with the relevance decreasing exponentially with a growth factor 0.5. The relevance of all the subjects is normalised so that the total relevance sums to one. The original paper doesn't mention the growth rate of their exponential dataset, making it difficult to replicate this dataset. We believe they used a growth factor of 0.5, due to the fact that they state that the distribution of the dataset and the geometric attention model are very similar and only differ in the number of postions with a value of zero, and that when we use this growth factor the results are very similar.

3.1.2 AirBnB datasets. The AirBnB data can be found online [6]. However, Biega et al. make no mention of the specific procedure they used to preprocess the data nor do they mention the specifics of the data. AirBnB's data is continually dumped, therefore we made an educated guess based on the publication date and the size of the datasets about which dataset they used. The datasets used for our research can be found in Table 1. However, since they made no mention of the preprocessing procedure we were unable to find the exact same amount of subjects.

Our preprocessing was done through regular expression finding the exact notions as specified by Biega et al.[7]. After removing the unnecessary columns, we dropped each row containing NaNs. This was done, firstly, for the multi-query database, and secondly, provided the basis for the single-query database. As such we could ensure the exact same size of all rankings. Table 1 display the sizes of the data we worked with.

3.1.3 StackExchange datasets. The StackExchange set was made from a specific datadump which wasn't clearly available online. Even with a different timestamp we had problems finding it. Biega et al. mentions that they take a subset from the data by excluding programming sub field but they make no mention of the exact list they exclude. This was then worsened by the fact that the specific preprocessing over this subset was subpar. We contacted Biega et al. to ask for the data but they gave no response. With the lack of clarity surrounding the specifics of the entire dataset and the lack of transparency in terms of preprocessing we decided to omit this from our reproduction.

[3] PuLP, https://pypi.org/project/PuLP/

[4] There are multiple solvers that can interface with PuLP, which one is used makes no difference in the final result.

[5] https://www.gurobi.com/

[6] www.insideairbnb.com

[7] The exact procedure was finding every column containing relevance dropping the rest (except for 'id') and then removing the final column, namely, *reviews_per_month*

3.2 Algorithm

3.2.1 Single- and Multi-Query. As in the original paper, we have conducted the experiment both in single- and multi-query fashion. We used the generated relevance scores to define the single-query within the synthetic dataset. Within the AirBnB dataset, we used the *review_scores_rating* attribute.

In the multi-query experiment, we used the following attributes from the AirBnB dataset to construct a sequence of rankings: *review_scores_rating*, *review_scores_accuracy*, *review_scores_cleanliness*, *review_scores_checkin*,
review_scores_communication, *review_scores_location*, and *review_scores_value.*

This approach is the same as in the original paper.

3.2.2 Singular and Geometric Attention. Regarding position bias, there are two approaches that were implemented by the original paper: singular and geometric attention. Singular means that only the first result receives attention, as is the case in some apps.

Geometric means that the attention has a fall-off depending on the rank and a certain cut-off point k. All positions after k have their weight set to 0. The parameters used for the geometric attention model are the same as in the original paper, which is $p = 0.5$ and $k = 5$. Naturally, the formula used to compute the attention value at position k is $w_k = (1 - p)^k p$.

3.2.3 Iterations. All experiments on the synthetic dataset were run for 300 iterations. The Airbnb dataset was run for 20.000 iterations on the single query and 3000 iterations on the 7-query (multi-query). This is exactly the same as the original paper.

3.2.4 Baselines. We use the same baselines the original paper provided, which allows allow for a better comparison between our results and theirs. The baselines used are:

- Relevance: ranking the results based on their relevance scores only.
- Objective: reranking the results without considering the quality of the ranking, purely based on increasing priority value. This priority value is calculated with the formula $A_i - (R_i + r_i)$, which makes it in line with the objective of minimizing unfairness.

4 RESULTS

In this section we go over the results and briefly discuss whether the hypotheses posed by Biega et al. hold. The main issue we had when interpreting the results by Biega et al., is that the provided figures do not clearly distinguish between different θ values and left us guessing in several cases. Hence, we added markers to allow for this distinction.

4.1 Results on the Synthetic dataset

We analyzed the model on the synthetic dataset as shown by Figure 1. On the x-axis we plot the iterations and on the y-axis our unfairness metric as described in Section 2.2, viewing these graphs we can see:

- For the uniform dataset with singular attention (Figure 1b), we see how the unfairness rises and then slowly decreases back to zero. Considering the algorithm this is of course fairly logical. Those who receive little attention become more preferable to the algorithm. Since all of them have the same relevance, we have to reach some kind of unstable equilibrium. Each cycle a new subject is placed at the top rank.
- For the linear dataset we can already see that there is a loss of quality when we introduce amortized fairness. Figure 1d can be found in the appendix.
- Regarding the linear dataset, we also experimented with smaller prefilter values (3 instead of 100, the same prefilter percentage, as in the Boston dataset). This experiment is visualised in Figure 3, and it shows that the fairness is negatively affected by the prefiltering.
- For the exponential dataset (Figure 1f and Figure 2f) we can see that the quality loss happens only when we lower the quality constraint beyond a certain θ value. Otherwise it will be the same as the relevance baseline.
- For geometric attention we can immediately see that the unfairness is decreased dramatically. Figures 2b 2d 2f can be found in the appendix. All of these findings are in line with the original paper.

4.2 Results on the Airbnb dataset

We analyzed the model on the Airbnb dataset which can be seen in figure 4 and figure 5. The axis are the same as the synthetic dataset. From these graphs we can conclude roughly the same things as the original paper. The trends are the same.

However these figures differ from the original paper in more ways than one. So the results are not perfectly reproduced but they do fall along the general hypothesis that unfairness is decreased when we introduce a kind of equity of attention.

5 DISCUSSION

It is far from practical that there is hardly any statistics and that the original results are mere graphs. This made it hard to compare our results to the original. Hence the comparison was based mostly on visual aspects of the graphs. While far from perfect it is the only option we have.

We contacted Biega specifically, to ask about the specifics of their Airbnb datasets, they did not respond, so the comparison here may be flawed on the basis that the datadumps are from different timestamps. This could explain the difference in relevances and the effect the reranking has on the fairness metric.

We quite accurately reproduced the results of the synthetic dataset. But beyond that exact reproduction has left something to be desired. There do not seem to be any glaring mistakes in our algorithm so most of the problems with reproduction could be boiled down to lack of clarity concerning the datasets. To counteract this we will make all our datasets publicly available.

An inclusion of values (such as unfairness after so many iterations) and an exact description of the dataset should improve the reproducibility tremendously. Furthermore Instead of NDCG we could have used Kendall's Tau. The original paper makes a mention of the possibility but never actually follows through, nor does it make any mention as to why they chose their implementation specifically.

Additionally, we argue that the trade-off that the authors describe as being between quality and unfairness is actually three dimensional, as the (speed) performance of the algorithm must be taken into account. The number of variables in the ILP problem is the prefilter amount squared, hence the speed of the algorithm largely depends on the prefilter amount. However, we also pointed out that this amount also greatly influences how the unfairness progresses over time. Biega et al. do not touch on the subject of performance, even though it took us a couple of hours to reproduce each one of their graphs on the AirBnB dataset (running several graph lines in parallel). It is not unimaginible that a company such as AirBnB receives thousands of requests per second. It would not be ideal if their users had to wait before the listing were displayed to them.

6 BROADER IMPLICATIONS

Does this algorithm actually improve fairness and how does it relate to the field of FACT? Well, the former still depends on your stance of fairness. The notion of proportionality and the notion of attention as a resource both seem like a reasonable improvement to the original model of ranking solely on relevance. With the introduction of attention as a resource we have a clearer representation of what we are redistributing. Of course this notion of fairness concerns itself with position bias, other biases based on sensitive features may still be present in the ranking. The algorithm presupposes that these are of lesser concern or at least keeps them out of the equation. Viewed through this lens, proportional attention as defined by Biega et al. may not be enough to ensure equal treatment under equal circumstances [14]. Furthermore, the algorithm makes no effort as to explaining why someone was preferred over another. The entire ranking itself is still opaque. We may argue that such an added explanation is beyond the scope of the paper but within this context we are introducing a more complicated model and this may reduce the modicum of understanding a layman can have about the ranker. Of course to a programmer who has access to the back-end we may see that the algorithm itself is rather transparent and easily auditable. Unlike the world of deep learning we aren't dealing with the same magnitude of opaqueness [15]. The last question concerning the field would be to look at the notion of confidentiality. How does this algorithm deal with the data provided? Is it safe? In essence it still uses relevance and thus there is no real change in the data used. We could ask whether such data should be privy to companies, but that is a question far beyond the scope of this paper. For that we need to turn to philosophy [16].

7 CONCLUSION

The notion of proportional attention in ranking systems does seem somewhat promising. It may not be very extravagant but nonetheless the notion introduces a clear idea about what resource they wish to redistribute within such systems. However, reproducing amortized fairness in ranking systems proved a bumpy ride. Some design choices weren't explained, the results were hard to compare and lacked statistical back-up. Overall the reproduction became a problem of interpreting instead of translating. This was worsened by the fact that the datasets lacked any clear description, which in turn made comparison neigh impossible. However, our results do show that fairness is accounted for and under certain constraints it can be "controlled" through the use of the algorithm. This is in line with their hypothesis. Without any artifacts to evaluate we can only award the original paper the Results reproducible badge.

Results Reproduced
acm

8 CONTRIBUTION

All members contributed equally. Emil and Nils focused more on the implementation and Thomas and Sietze focused more on the paper.

REFERENCES

[1] M. Heidegger. *Discourse on Thinking*. Harper Perennial Modern Thought. HarperCollins, 1969. ISBN 9780061314599. URL https://books.google.nl/books?id=GFIW9D8Z0c0C.

[2] Tefko Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance. *Journal of the American Society for information Science and Technology*, 58(13):2126–2144, 2007.

[3] Lucas D Introna and Helen Nissenbaum. Shaping the web: Why the politics of search engines matters. *The information society*, 16(3):169–185, 2000.

[4] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3150–3158, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/liu18c.html.

[5] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NIPS Tutorial*, 2017.

[7] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. *arXiv preprint arXiv:1712.03586*, 2017.

[8] Robert Nozick. Distributive justice. *Philosophy & Public Affairs*, pages 45–126, 1973.

[9] Thomas Scanlon. Nozick on rights, liberty, and property. *Philosophy & Public Affairs*, pages 3–25, 1976.

[10] John Rawls. *A theory of justice*. Harvard university press, 2009.

[11] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 405–414. ACM, 2018.

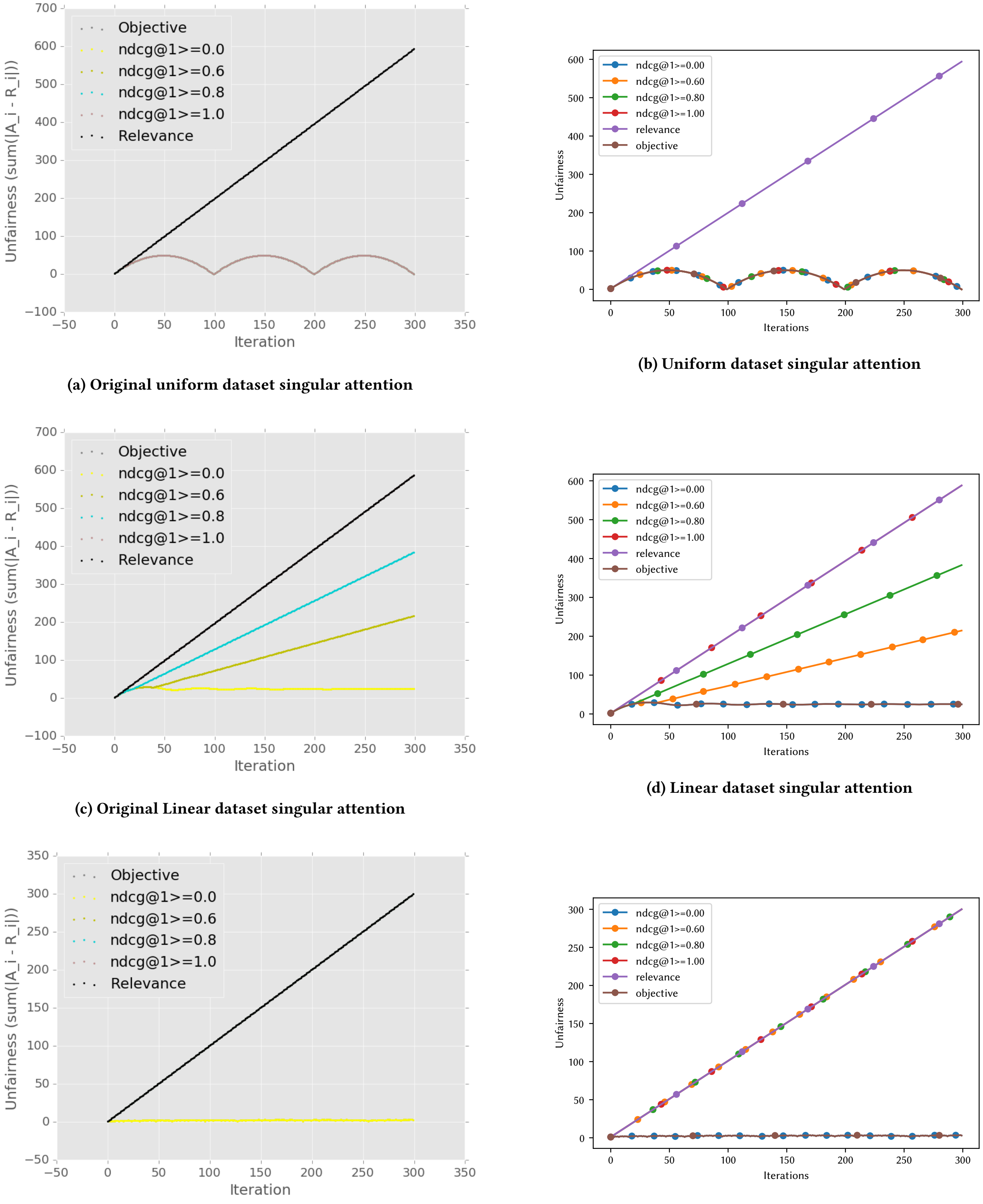
[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[13] Jakob Nielsen. F-shaped pattern for reading web content (2006). *Im Internet: www. nngroup. com/articles/f-shaped-pattern-reading-web-content*, 2007.

[14] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

[15] José Mena, Oriol Pujol, and Jordi Vitrià. Dirichlet uncertainty wrappers for actionable algorithm accuracy accountability and auditability. *arXiv preprint arXiv:1912.12628*, 2019.

[16] Michel Foucault. *Discipline and punish: The birth of the prison*. Vintage, 2012.



(a) Original uniform dataset singular attention

(b) Uniform dataset singular attention

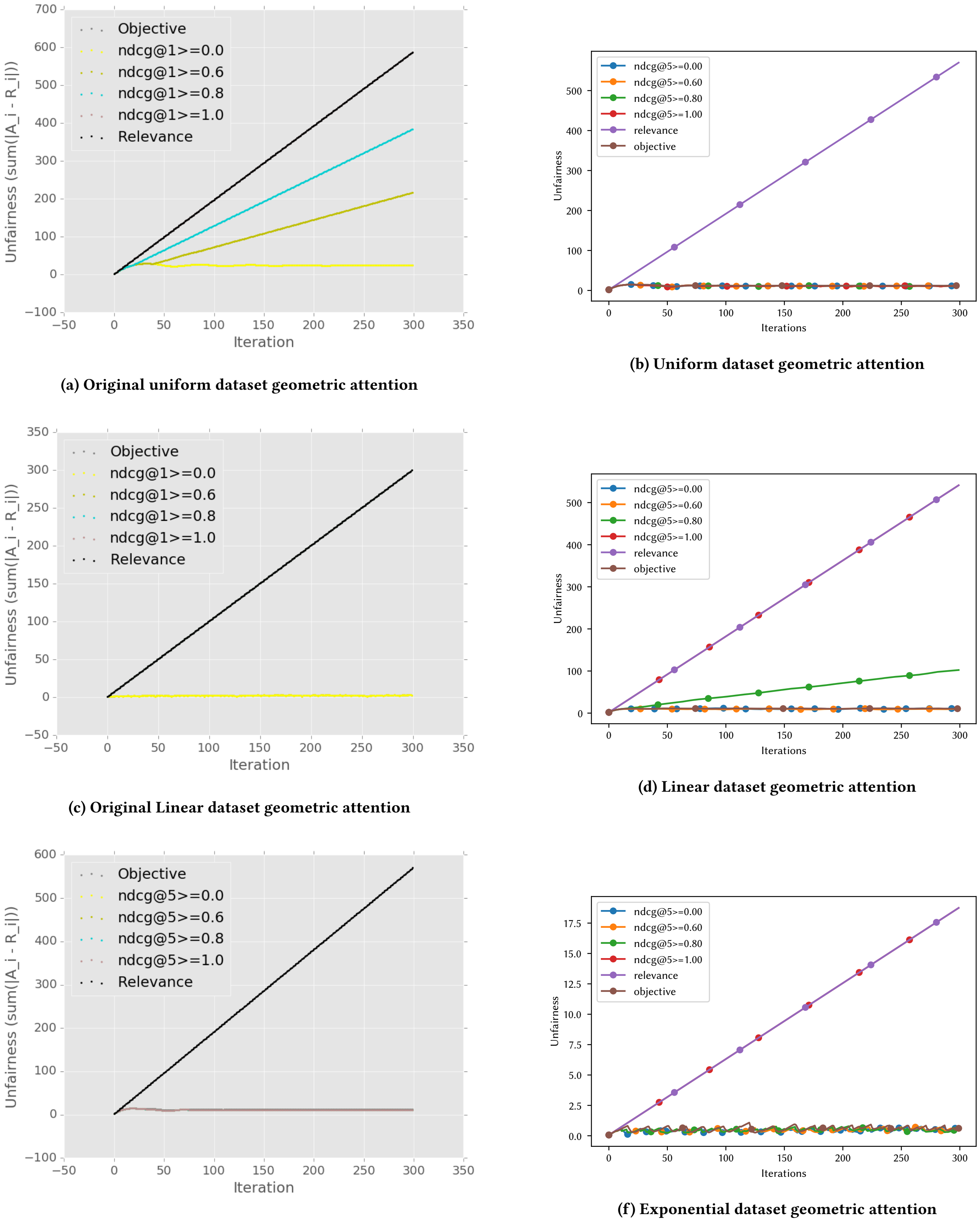
(c) Original Linear dataset singular attention

(d) Linear dataset singular attention

(e) Original Exponential dataset singular attention

(f) Exponential dataset singular attention

Figure 1: Synthetic dataset singular attention



(a) Original uniform dataset geometric attention

(b) Uniform dataset geometric attention

(c) Original Linear dataset geometric attention

(d) Linear dataset geometric attention

(e) Original Exponential dataset geometric attention

(f) Exponential dataset geometric attention

Figure 2: Synthetic dataset geometric attention

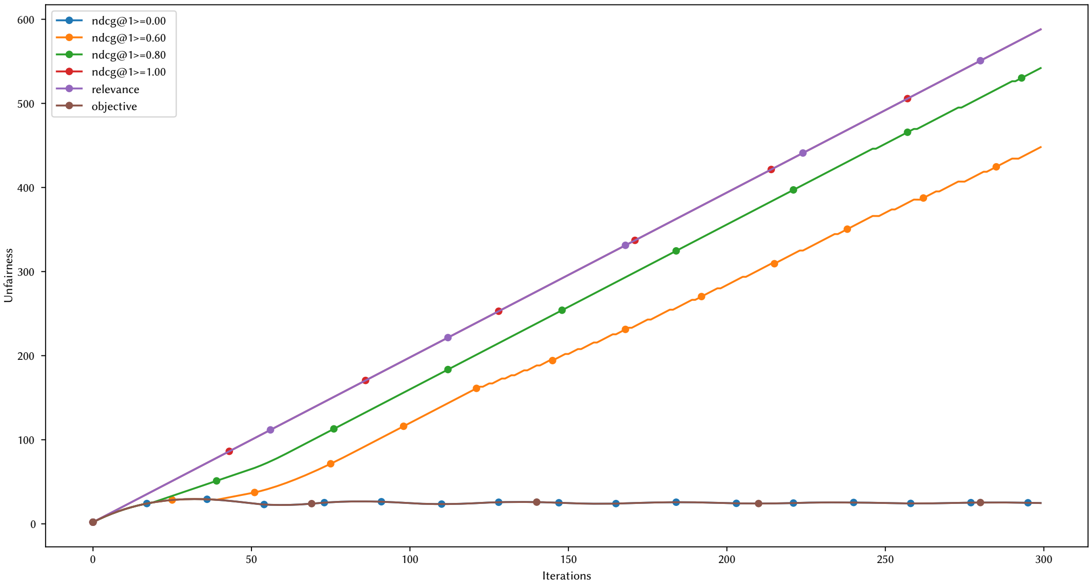
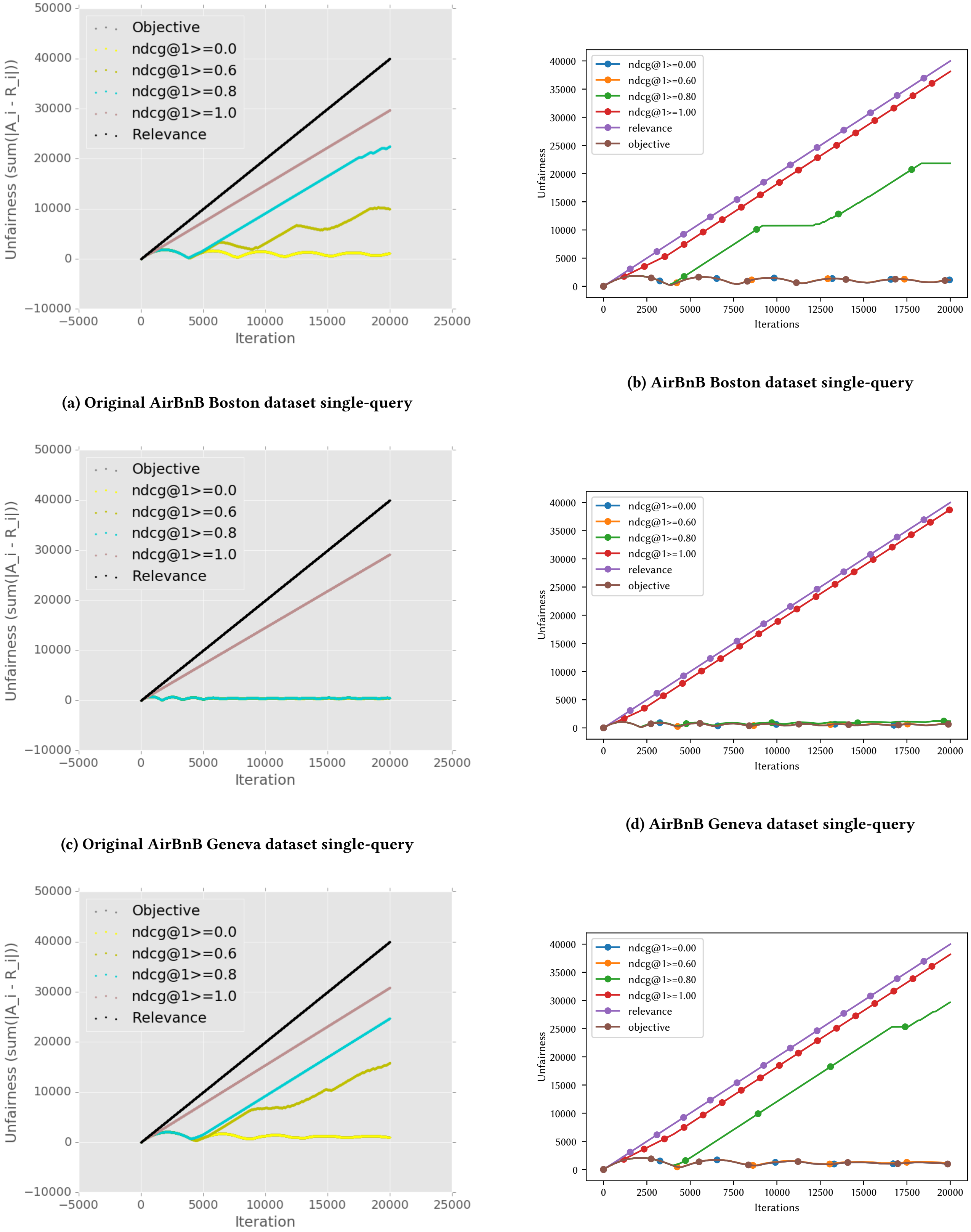


Figure 3: Linear dataset singular attention, small prefilter



(a) Original AirBnB Boston dataset single-query

(b) AirBnB Boston dataset single-query

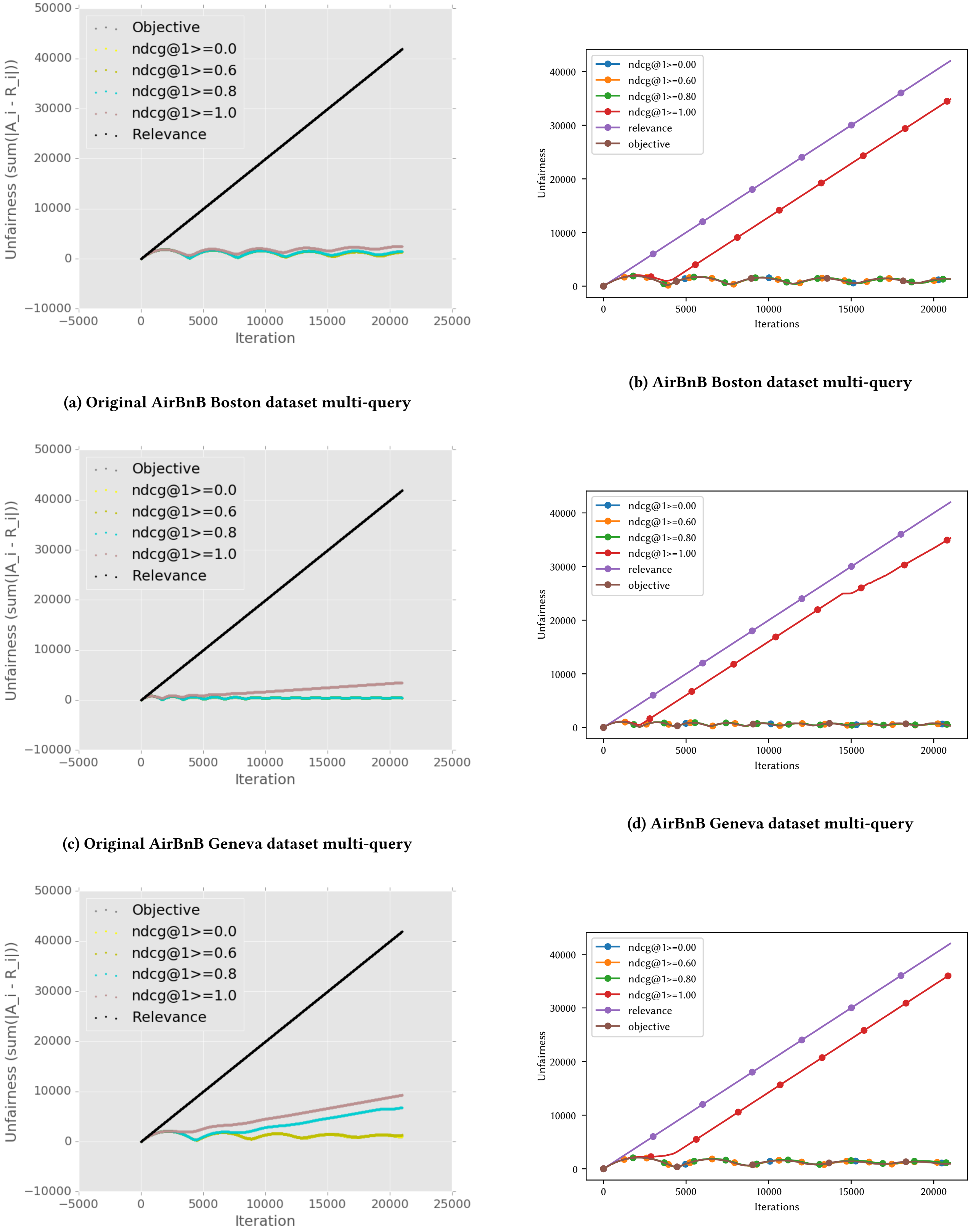
(c) Original AirBnB Geneva dataset single-query

(d) AirBnB Geneva dataset single-query

(e) Original AirBnB Hong Kong dataset single-query

(f) AirBnB Hong Kong dataset single-query

Figure 4: Single-query AirBnB dataset



(a) Original AirBnB Boston dataset multi-query

(b) AirBnB Boston dataset multi-query

(c) Original AirBnB Geneva dataset multi-query

(d) AirBnB Geneva dataset multi-query

(e) Original AirBnB Hong Kong dataset multi-query

(f) AirBnB Hong Kong dataset multi-query

Figure 5: multi-query AirBnB dataset