Master's Degree in Specialized Economic Analysis

The Implications of Algorithms in Rental Markets with Matching Mechanisms. The Role of Exposure, Choice and Efficiency in the Sharing Economy.

Authors: Nikhil Bhutada, Marc Miró i Escolà, and Daniel Müller-Demary

Directors: Ada Ferrer i Carbonell and Hannes Mueller

June 2018

ABSTRACT IN ENGLISH

Given the increased use and dependence of peer-to-peer markets on algorithmic optimisation processes, the problems biased algorithms may bring about for individuals have been gaining significance. With the emergence of the sharing economy, more people can be subjected to damaging allocation systems. This paper makes a first attempt to connect the notion of bias, frequently discussed in computer science, with classic problems of systematic discrimination observed in labour markets. Following the methods used in identifying such biases, the authors discuss aspects of equity and efficiency that arise from different matching mechanisms. In offering a framework for analysing the algorithms present in such markets, the authors conclude that any verdict depends on the normative stance of the reviewer, leaving the reader with a set of questions and directions for future research.

ABSTRACT IN CATALAN

Donat l'increment de l'ús i la dependència dels mercats peer-to-peer (mercats entre pars) de processos d'optimització algorítmica, els problemes que poden ocasionar els algoritmes esbiaixats per als individus han estat guanyant importància. Amb l'emergència de l'economia col·laborativa, més gent pot estar subjecte a sistemes d'assignació perjudicial. Aquest article fa un primer intent de connectar la noció de biaix, freqüentment discutit en la ciència informàtica, amb els problemes clàssics de discriminació sistemàtica observats als mercats de treball. Seguint els mètodes utilitzats en identificar aquests biaixos, els autors discuteixen aspectes com l'equitat i l'eficiència que sorgeixen de diferents mecanismes d'emparellament. En oferir un marc per a analitzar els algoritmes presents en aquests mercats, els autors conclouen que qualsevol veredicte depèn de la posició normativa del revisor, deixant el lector amb una sèrie de preguntes i direccions per a futura recerca.

KEYWORDS IN ENGLISH:

Peer-to-Peer Markets, Sharing Economy, Algorithm Bias, Eficiency, Equity, Matching Theory

KEYWORDS IN CATALAN: Mercats entre pars, economia col·laborativa,

biaix d'algoritme, eficiència, equitat, teoria d'emparellament

The Implications of Algorithms in Rental Markets with Matching Mechanisms. The Role of Exposure, Choice and Efficiency in the Sharing Economy.

Daniel Müller-Demary[*] Marc Miró i Escolà [†] Nikhil Bhutada[‡][§]

12 June 2018

Abstract

Given the increased use and dependence of peer-to-peer markets on algorithmic optimisation processes, the problems biased algorithms may bring about for individuals have been gaining significance. With the emergence of the sharing economy, more people can be subjected to damaging allocation systems. This paper makes a first attempt to connect the notion of bias, frequently discussed in computer science, with classic problems of systematic discrimination observed in labour markets. Following the methods used in identifying such biases, the authors discuss aspects of equity and efficiency that arise from different matching mechanisms. In offering a framework for analysing the algorithms present in such markets, the authors conclude that any verdict depends on the normative stance of the reviewer, leaving the reader with a set of questions and directions for future research.

Keywords: Peer-to-Peer Markets, Sharing Economy, Algorithm Bias, Efficiency, Equity, Matching Theory
JEL: H21, D63, C78

[*] Barcelona Graduate School of Economics. daniel.demary@barcelonagse.eu

[†] Barcelona Graduate School of Economics. marc.miroescola@barcelonagse.eu

[‡] Barcelona Graduate School of Economics. nikhil.bhutada@barcelonagse.eu

[§] The authors would like to take this opportunity to thank the faculty at Barcelona Graduate School of Economics with special thanks to Hannes Mueller, Ada Ferrer i Carbonell, Caterina Calsamiglia and Augustin Tapsoba. Furthermore, we would also like to thank Badi and especially Guillem Pons for providing us with the dataset for the thesis.

1 Introduction

The sharing economy is a network of peer-to-peer based activities of giving or obtaining access to goods and services through community-based online media (Albinsson and Yasanthi Perera, 2012). With respects to the rental market, the sharing economy can be defined as private individuals renting property to each other. This kind of activity has experienced significant growth in the last decades as the channels through which suppliers and buyers find each other have changed. Through the advent of apps and digital platforms mimicking bulletin boards, markets previously inaccessible to individuals due to time and informational constraints have relaxed their barriers of participation (Hamari et al., 2016; Einav et al., 2016). By charging a commission on each transaction, the companies providing the exposure that individuals require for participating in such sharing economies, have grown alongside the markets they facilitate.

Typically, markets based on the sharing of property, as is the case in housing or carpooling, follow the design of a two-sided matching market, in which both the proprietor as well as the person demanding the service must explicitly agree on the transaction beforehand. The additional layer of mutual approval may stem from the heterogeneous quality of the service provided or goods rented, in which case the individuals seeking to rent distinguish between offers made by attractive and unattractive service providers. For instance, in the context of car sharing, this may be observed in the renter's choice being subject to reputation checks (e.g. Uber's star system) (Rauch and Schleicher, 2015).

The need for transaction approval can be equally important to the service provider: a person renting his room to a stranger will first ensure themselves that the potential tenant fits their expectations. The importance and existence of such confirmation-mechanisms are ultimately subject to individual market properties (i.e. the type of good exchanged), nevertheless retaining a central role in the shared economies that have flourished between private individuals.

What sets these markets apart from those typically discussed in the literature on matching is the continued use of the price as a market mechanism. Classically, markets relying on matching do not incorporate features such as pricing because matching arises from the constraints imposed by regulation, as in the case of kidney exchanges (Roth et al., 2004) or because the efficiency resulting from the price mechanism implies the optimal allocation of goods, rendering other mechanisms as futile (Fama, 1970; Debreu and Scarf, 1963; Malkiel, 2003).

The matching literature has mostly dealt with scenarios that assess the efficiency of allocation in the absence of prices (Roth, 1982; Shapley and Scarf, 1974; Roth and Sotomayor, 1992; Niederle et al., 2008) identifying measures and conditions such as thickness, congestion, stability or strategy-proofness that play a key role in achieving efficient allocation. Yet the prevailing algorithms (e.g. Top Trading Cycles) do in turn, rest on assumptions regarding the preferences of agents.

This paper discusses a case of the sharing economy that acts as a "hybrid market" relying on prices for allocation but also including a condition of mutual confirmation that introduces a type of matching procedure. It is worth noting that the case discussed here violates some of the underlying assumptions adopted in the classic literature. The authors first contextualize the existence of such hybrid markets, identifying the conditions under which they appear, then discuss the current approaches to optimizing allocation and the problems that may arise from such methods. The paper reviews the matching procedures used in such markets and the ramifications they present, and then proposes a method that allows for the identification of matching-related problems in more general conditions. To conclude, we discuss the distributional outcomes of matching mechanisms and comment on directions for improvement.

1.1 The market design in the sharing economy

In the context of the service-based sharing economy, individuals can be categorised as either *seekers* who are looking to purchase a service, or *listers* which are offering their services on the marketplace. Due to the personal nature of the transaction, the manner in which the service is enjoyed by both parties and the satisfaction of individuals after the completion of the service plays a defining role in the utilities of the lister and the seeker. The lister attempts to maximize their utility by offering their service at the market price to a seeker whose consumption style will not produce any negative externalities (as is the case for renting rooms to noisy tenants) nor affect the quality of the good subsequently (e.g. the furniture in the rented space should not suffer damages resulting from the stay). Conversely, the seeker's utility is also defined by the quality and limitations of the service and the price required.

The existence of different types of seekers leads to a signaling game, in which individuals looking to make use of a service are required to present themselves in a favourable light to the providers. The diversity in the quality of the services provided implies that the listers must also present themselves accordingly, to appeal to the right seekers. The presence of these factors leads to a compatibility check which can be seen as a practical implementation of a matching mechanism, subject to the same characteristics as those discussed in the literature: Thickness, congestion, stability or strategy-proofness (Niederle et al., 2008).

1. Thickness is the property of a market to provide enough listers and seekers to each other as to make participation in the market efficient. Comparable to a state of competition leading to efficient prices, a thick market leads to pairs of individuals that are more efficient.

2. Congestion is a consequence of mismanaged thickness, in which the lack of sufficient time for individuals to compare all available offers leads to hastened, suboptimal decisions. Due to time constraints, congestion is often identified in sharing economies (Fradkin, 2017).

3. Stability is the property of a match rising from the fact that neither individual in the pair can find another outside that they would prefer and in turn would be preferred by. In other words, a pair of a seeker and a lister is stable if neither can find another better partner to match with that would also prefer them to their own match (Gale and Shapley, 1962).

4. Strategy-proofness is an emergent property of the mechanism of matching by which individuals experience no profitable deviation in their matches by misrepresenting their preferences or qualities. Strategy-proofness is a characteristic of the algorithm used to group individuals into pairs that has widespread consequences for the efficiency of the matches (Dutta and Massó, 1997).

1.2 The Case of a Room-Renting Market

The market discussed in this paper is one of room rentals in the city of Barcelona. The listers and the seekers are individuals who own or would like to rent rooms in the city and interact with each other through the application or online platform developed by a company named Badi [1]. Due to the large number of individuals active on the platform, it is impossible to identify the entire set of preferences a user has for potential matches (it is unfeasible to ask a user to rank every lister or seeker), leading to a state of congestion as people seek to transact faster than efficient matching would allow, as well as a lack of stability stemming from the influx of new users. To constrain matters more, the demand for individuals bearing specific characteristics leads to conditions where seekers and listers mask their own attributes for more desirable ones in an attempt to match with more individuals (e.g. silent roommates are preferred, so people will present themselves as such).

[1] https://www.badi.com

Companies like Badi offering to improve matching in such markets therefore rely on algorithms that operate in spite of the constraints imposed. Taking into account the limited attention span that users have when rating potential partners, the company's algorithm attempts to maximize the number of matches by suggesting individuals to each other which are more likely to be preferred. Instead of asking users to rank all potential matches, the platform reduces this to a binary decision: "Do you consider this person a suitable potential flatmate?". If both users opt "yes" they may start a chat and proceed further. The algorithm minimizes the length of the search process a user undergoes to find a good match, by reducing the number of partners users have to sift through. A consequence of such an approach is the fact that an individual's exposure to the market, and therefore the quality of the service provided, is now contingent on how suitable the algorithm considers them for matching. If a relatively unpopular individual that had a low but nonzero probability of finding a room, has an even lower chance under the algorithm, one may consider this loss as indicative of discrimination. Consider for instance, the possibility of an algorithm that systematically diminishes the exposure of a specific gender or nationality. The existence of such algorithms has been subject to intense discussion, predominantly in the literature of machine learning and computational statistics under the term *algorithmic bias* (Crawford et al., 2014; Kirkpatrick, 2016; Hajian et al., 2016; Richterich, 2018).

Instances of discrimination or welfare loss resulting from poorly developed algorithms have been found in policing systems and judicial institutions (Angwin et al., 2016), in commercial access to services (Crawford, 2016), as well as in differential treatment in online representation (Sweeney, 2013). In recognizing the proliferation of algorithmic approaches in the shared economy, the possibility of systematic discrimination bears more weight in the design of algorithms and matching markets than ever before. In the following section we outline an approach that identifies such biases.

2 Method

2.1 Matching Procedure and Data Collection by Badi

Prior to identifying patterns of discrimination in the market hosted by Badi, we explain the procedure that individuals undergo to find a room or roommate. When an individual seeks access to the market through the platform, they must first specify whether they are a seeker or a lister. Following this classification, they are asked to create a profile in which they state their name, age, gender, place of residence, short biography, and personality. The personality is described with the help of six scales ranging from zero to ten, where a larger number represents a greater inclination for a trait. The personality traits measured are: degree of sociability, athletic inclination, degree of tidiness and organisation, degree of “geek” culture, openness to party, and the willingness to partake in an active lifestyle.

As the ratio of seekers to listers is five to one, the company aims to maximize the number of matches between individuals by first targeting the preferences of the listers. The procedure for matching is as follows:

1. Multiple seekers are first suggested to each lister in their local area. This is done without specific actions from either user, through a randomised procedure in which all seekers have the same probability of getting suggested (conditional on certain priors), or through the medium of the algorithm which suggests certain seekers to specific listers.

2. Following the suggestions, the lister responds: “Yes”/”No”. If and only if the lister’s response is positive, a message is sent to the seeker notifying him of a potential flatmate. The relationship between a seeker and a lister at this stage is called a *Semi Match*, as the lister is interested, but the seeker has not yet reciprocated.

3. If the seeker responds in kind, a *Final Match* is created, which opens up the the possibility to chat with each other. Figure 1 illustrates this in more detail below.

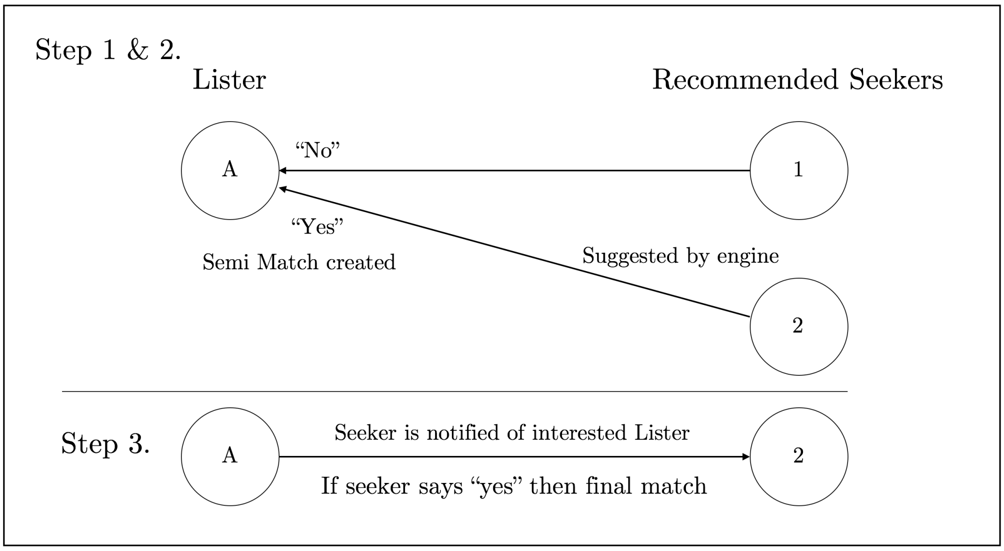

Figure 1: Diagram of the matching process in Badi

Both seekers and listers may procure an unlimited number of final and semi matches. Both the random and the algorithmic recommendation engines will only stop suggesting a seeker profile after two weeks of inactivity.

The relationship between the nature of the suggestions and the amount of semi and final matches may be modeled as follows: the recommendation engine may change the seeker's *exposure* by varying the amount of times they are suggested to listers. Additionally, the *quality* of the suggestions may change, that is, the seeker is suggested to listers that are more likely to respond positively to their profile. While the change in quality is an implicit property of the recommendation engine, the exposure is directly observable through the number of times a seeker is suggested. Figure 2 illustrates this below. We hypothesize that both these parameters affect the number of semi matches a seeker experiences while active on the platform.

Following a review of the determinants of semi matches, we explore the changes in final matches a seeker experiences as a result of being subjected to the algorithm. Once more, the recommendation engine may affect the number of final matches in two ways: by increasing the amount of listers a seeker can choose from (effectively boosting the number of semi matches) or again through the channel of recommendation quality. Such a case in which a seeker receives the same amount of interested listers under the algorithm as under the random engine, yet has higher number of final matches would be due to the channel of quality. Having established all the possible channels through which the recommendation engine may affect the quality of the matching service an individual relies on, we turn our attention to the prospect of algorithmic bias.

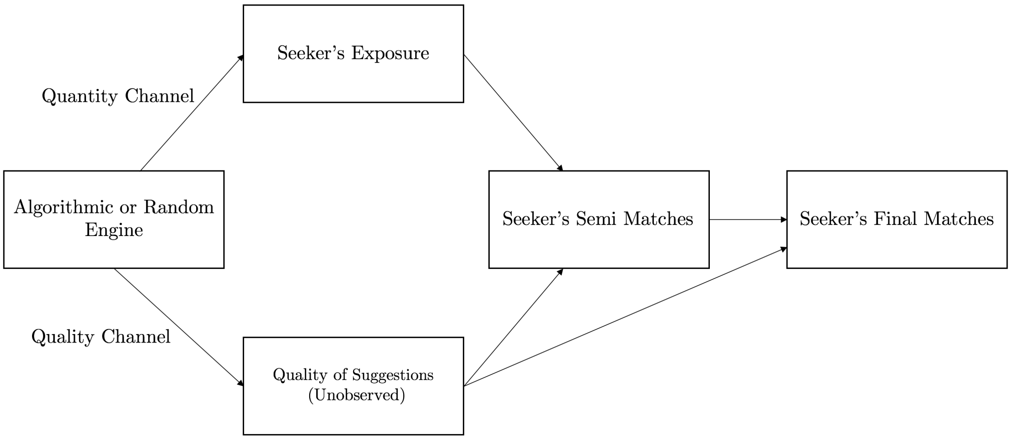

Figure 2: Relationship between exposure, semi matches and final matches

2.2 Algorithmic Bias

Algorithmic Bias is the systematic discrimination of subgroups through a negative change in the services provided. As pointed by Hajian et al. (2016), such biases may exist in the absence of bad intentions from the platform developers. This may result from algorithms being trained on an unrepresentative dataset which in turn incorporates the harmful connections (for a detailed case see Angwin et al. (2016)). But even a company suppressing sensitive variables from the inputs of their matching service in hopes of sidestepping the problem, may run into complications. Discrimination can arise from correlations in the data, where the omitted variables are grouped simply through their association to other trends (Hajian et al., 2016).

Given the fact that listers and seekers are subjected to slightly different procedures, this paper focuses its effort on the seekers as they make up the vast majority of the population. The listers may also experience bias however given the high demand for rooms we recognize this to be of little importance. In the context of the matching market discussed here, a bias would entail a significant decrease in the number of semi matches an individual amasses. This definition hinges on the objective function of the algorithm - to maximise the number of semi matches. However, it is worth considering at what cost a seeker receives interested listers.

As mentioned before, the attention of listers is quasi-limited, meaning the task of suggesting seekers to them takes the form of a zero sum game. While no formal constraint exists, it can be assumed that seekers compete for the attention of the listers. As such, one must consider how many resources (how many listers should be informed) have been allocated to achieve a number of semi matches for a seeker. The algorithm may therefore change the experience of the seeker (with respect to the random engine) in a variety of ways, as shown in figure 3 and explained below.

A matching algorithm may present no visible advantage at creating semi matches for a specific seeker. As is shown in the upper right quadrant in figure 3, any increase in the number of semi matches is a direct result of a specific seeker being recommended more. This setting may be considered typical; the advantage provided by the algorithm versus the random engine rests on the fact that a user receives more exposure. However such an approach cannot be sustainable for all users, due to the soft zero-sum constraint explained above. If some users gain attention through more suggestions, others will lose some. A well designed algorithm may address this constraint and recommend seekers to a smaller, but more suitable group of listers. This may be regarded as the ideal scenario (upper left quadrant); less attentional resources are used per seeker to achieve a higher number of semi matches.

The relationship between exposure and semi matches may in some cases even be negative and would stem from a badly designed algorithm. An

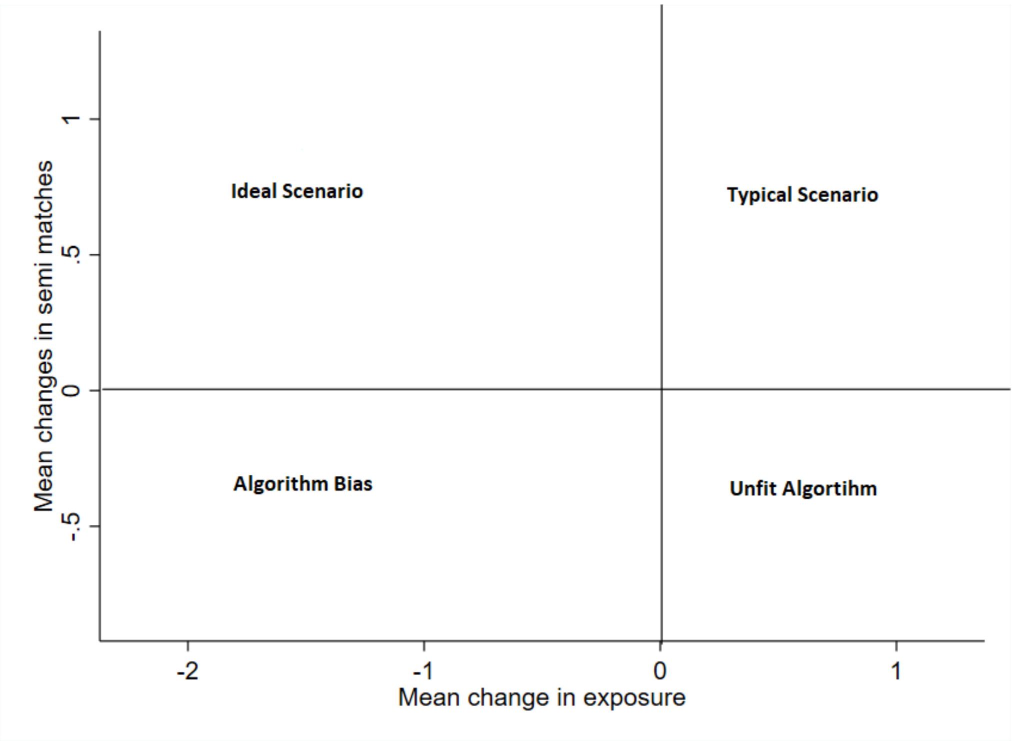

Figure 3: Potential scenarios

algorithm that requires more resources and produces less output than a random engine is clearly pareto-inferior (lower right quadrant, "Unfit Algorithm").

The final scenario we consider is one of algorithmic bias. An algorithm that decreases the exposure a seeker receives as well as the number of semi matches could be said to intentionally marginalize a person from the marketplace, effectively discriminating. The choice of these two criteria in defining bias (decrease in exposure and semi matches) rests on both normative and legal arguments: services discriminating on the basis of gender, ethnicity or health status are prohibited by numerous legal systems (e.g. the Spanish constitution[2]) as well as the universal declaration of human rights[3].

By making use of this technique, one can visualise the change in the relationship between exposure and semi matches experienced by each individual seeker in figure 4.

[2] art.14 CE, 1975

[3] UN G. A. 1948

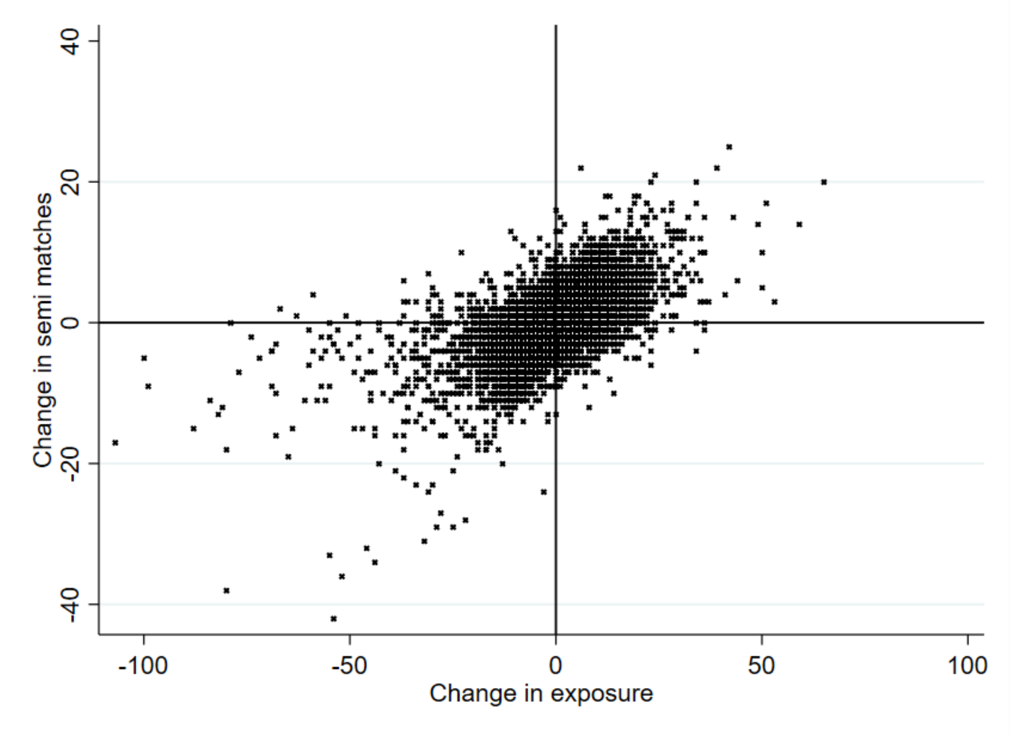

Figure 4: Individual level bias

As is visible in the figure, a substantial amount of individuals do seem to fall victim to what we have defined as bias (lower left quadrant). Yet it remains unclear at what point this bias starts or where one may set the threshold (How far does one have to deviate from the origin?). As the data depicted is measured at an individual level, there is no possibility for inferential statistics, therefore a method relying on intervals must be implemented. Moreover, algorithmic discrimination is problematic when it *systematically* targets subgroups of the population on the basis of gender, age or ethnicity. In figure 4, it is not clear if the seekers have fallen victim on the basis of such criteria. We introduce a method to identify the existence of biases at group levels. A further section is dedicated to the relationship between changes in semi matches and those in final matches in a discussion around efficiency, utility and welfare.

2.3 Group Level Discrimination

In order to establish the existence of group level biases one must identify consistent discrimination for a specific subgroup. While conditioning on age, nationality and gender may appear to be a sensible approach, consider the interactions of these traits with personality characteristics. The algorithm may work differently for seekers with different personalities, and the distribution of gender, age and nationality over the different personality types presented itself is highly uneven. Identification therefore took the following form:

$$\begin{aligned}\text{Bias}_{seeker\ group} = &(\#Semi\ matches|Age\ group, gender, personality, random)\\ &-(\#Semi\ matches|Age\ group, gender, personality, algorithm)\\ &conditional\ on\ lower\ exposure\ in\ the\ algorithmic\ case.\end{aligned}$$

A positive significant difference indicates a bias.

The dataset offered by Badi introduced some general limitations: the vast majority of users were Spanish (approximately 95%) meaning that our analysis could not be conducted on the basis of nationality, due to insufficient data. Restricting the evaluation to age and gender however led to a second limitation. As seekers rated their personality on six dimensions each consisting of eleven increments (0-10), the number of cases one had to condition on to compare became unfeasibly large (gender:2, age:50+, personality type: 11ˆ6). To address the problem of dimensionality, we took advantage of the k-means clustering algorithm (MacQueen et al., 1967) and used the elbow method (Goutte et al., 1999; Makles, 2012) to identify the suitable number of groups. The elbow method relies on calculating the sum of squared errors within clusters (WSS) and overall (TSS), for a given number of groups. By computing the ratio of the WSS to TSS, known as η^2 for multiple clusters k, one can trace the rate at which variance becomes explained by the groups.

$$\eta_k^2 = 1 - \frac{WSS(k)}{WSS(1)} = 1 - \frac{WSS(k)}{TSS} \forall k \in K$$

The optimal number of clusters is found at the last point on the plot at which the increase in variance explained by the groups is visibly larger, after which splitting the population into even more clusters does not add any information but merely leads to overfitting. The data suggests that splitting the seekers into seven groups on the basis of the six personality dimensions is optimal. The elbow plot is in the appendix (Figure 7), while the table below presents a descriptive label for each group (see appendix for means of each characteristic for each cluster in Table 9).

Cluster	Defining Property
Cluster1	Introvert and disorganized
Cluster2	Extrovert and not athletic
Cluster3	Lazy organized
Cluster4	Partying geek
Cluster5	Social organized
Cluster6	Lazy geek
Cluster7	Disorganized lonely

The other variables that had their dimensions reduced were seeker and lister age. The population sizes at different ages vary slightly between the listers and the seekers because the mean age for the listers (34) is higher than that of the seekers (30). The age intervals are: 1 (18-23), 2 (24-29), 3 (30-40) and 4 (over 40). For both cases, the proportion of listers and seekers in each age group is roughly equal. Having addressed the disparity between the number of seekers in our dataset (43,747) and the number of dimensions (≈10ˆ9) we sorted our seekers into 56 groups (sex:2×age group:4×personality cluster: 7) and applied the previously explained method on bias identification, leading to the following graph.

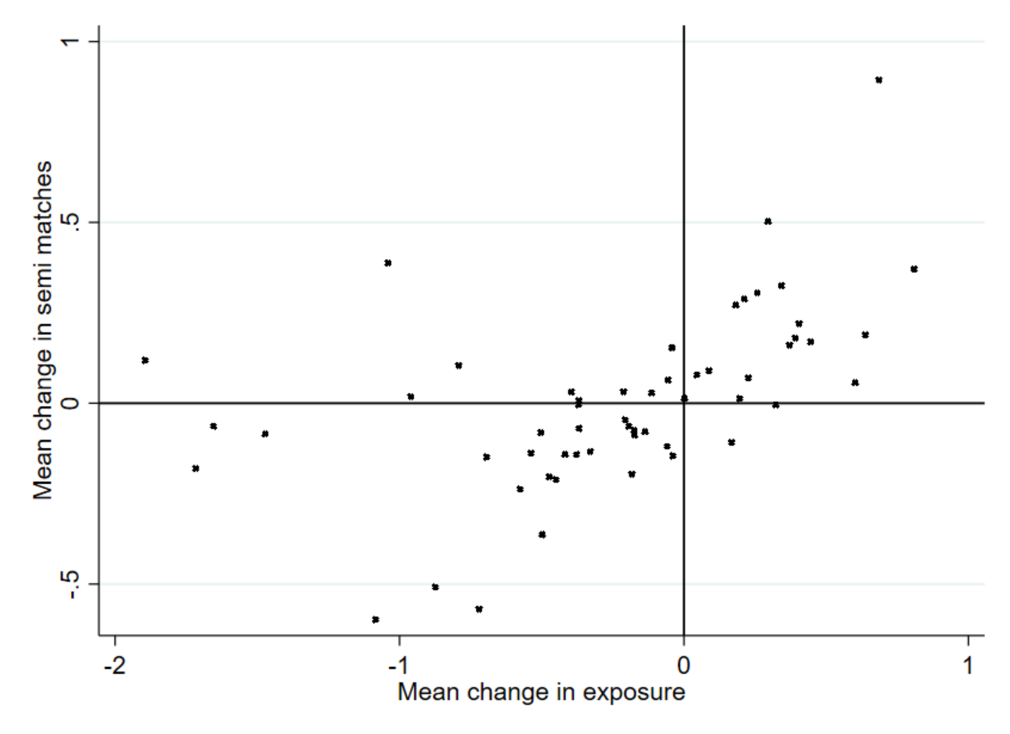

Figure 5: Group level bias

In the graph above at the group level, the majority of observations do fall into one of the four quadrants. The added advantage of a group level analysis is the introduction of significance thresholds. While the majority of the points do not intersect the axes, hypothesis tests conclude that only four cases experienced significant changes in semi matches as well as exposure, when transitioning from random to algorithmic. The tables below review these differences.

2.4 Summary

Table 1 outlines the fact that the amount by which females are recommended through the algorithm or the random engine does not diverge significantly for most groups. Only for the three cases of the second cluster (extrovert and non-athletic) in the age intervals 18 to 40, is the algorithm exposing the seekers less (sig. at 10%). In table 2, we observe that the lower exposure of the three groups does not translate into significantly fewer semi matches. While we cannot comment on whether the algorithm is truly more resource-efficient, we can state with certainty that females are not subject to bias.

Table 1: Average exposure for each female

		Clusters						
Age		**1**	**2**	**3**	**4**	**5**	**6**	**7**
	Algo	4.5680	4.7959	5.0446	5.1130	6.0197	5.1766	4.5125
18-23	Random	4.4280	5.1980	5.2030	5.1815	6.2988	5.6632	4.8172
	p-value	0.5255	0.0765	0.5121	0.8763	.5639	0.2415	0.5464
	Algo	4.8376	4.9495	5.1673	5.2973	5.2393	5.8936	5.6028
24-29	Random	4.8076	5.3726	5.2658	5.3198	5.6312	6.2284	5.7399
	p-value	0.8953	0.0260	0.6161	0.9522	0.1206	0.4306	0.6856
	Algo	4.9390	5.7679	5.8252	7.4644	5.9553	6.0572	5.8598
30-40	Random	5.0971	6.4270	6.1313	8.1423	6.4262	6.6114	5.9800
	p-value	0.5098	0.0862	0.3393	0.4858	0.2055	0.3570	0.7771
	Algo	4.9478	6.0096	5.7674	4.8182	5.9636	5.8889	6.1437
41-77	Random	5.2609	7.0192	6.0242	4	5.8461	7.2889	6.5375
	p-value	0.5435	0.4492	0.6283	0.3942	0.8391	0.2573	0.6119

For the male seekers, table 3 shows that for two groups (cluster 1 ages 30-40; cluster 5 ages 24-29), the exposure increases significantly under the algorithm. In the table 4, cluster 6 (lazy geeks) aged 24 to 40 has benefited from the algorithm with increased numbers of semi matches. These cases may be considered "ideal scenario" as they achieve a favourable outcome with the same exposure. Additionally, individuals in cluster 5 (social organized) aged 18-23, 30-40 and 41-77 and cluster 4 (partying geek) aged 24-29 have also benefited from the algorithm. Overall, no group-specific bias has been identified. Results indicate that the algorithm is performing

Table 2: Mean semi matches for each female cluster

					Clusters			
Age		**1**	**2**	**3**	**4**	**5**	**6**	**7**
18-23	Algo	0.3474	0.3781	0.3784	0.3781	0.3682	0. 3474	0.4006
	Random	0.3673	0.3861	0.3791	0.3554	0.3632	0.3559	0.3682
	p-value	0.2625	0.5540	0.9610	0.3970	0.8349	0.7008	0.2325
24-29	Algo	0.3860	0.4108	0.3870	0.3811	0.4055	0.3944	0.3879
	Random	0.3872	0.4124	0.3896	0.4161	0.3939	0.3743	0.3889
	p-value	0.9358	0.8887	0.8056	0.1008	0.4231	0.2668	0.9549
30-40	Algo	0.3298	0.3733	0.3699	0.3777	0.3959	0.3168	0.3614
	Random	0.3410	0.3672	0.3560	0.3924	0.3869	0.3924	0.3851
	p-value	0.4770	0.7201	0.4220	0.6278	0.5719	0.0026	0.2276
41-77	Algo	0.2331	0.3082	0.2646	0.4442	0.2743	0.2670	0.2359
	Random	0.2267	0.2560	0.2783	0.1818	0.2801	0.2424	0.2472
	p-value	0.7831	0.2501	0.4442	0.0053	0.8359	0.6869	0.7416

slightly better for the males than for the females. Amongst the males, the "social organized" and "lazy geeks" have benefited the most, while amongst females it is the "extrovert and not athletic" group.

Table 3: Average exposure for each male

					Clusters			
Age		**1**	**2**	**3**	**4**	**5**	**6**	**7**
18-23	Algo	4.5976	5.2729	5.5631	5.8436	5.7965	5.4850	5.6071
	Random	4.7422	5.2439	5.5072	5.8651	5.6216	5.7462	5.8801
	p-value	.6913	.9288	0.9026	0.9558	0.6949	0.4931	0.6492
24-29	Algo	5.4072	6.0012	5.9380	5.9621	6.4238	6.1513	6.1040
	Random	5.2002	5.6703	5.5427	5.6722	5.9247	6.0686	5.8768
	p-value	0.5089	0.1956	0.1705	.2661	0.0399	.7823	0.5558
30-40	Algo	6.1717	7.0943	6.8640	6.8683	7.3048	7.1623	6.8876
	Random	5.5327	6.8306	6.8026	6.6695	7.0877	7.0030	6.7609
	p-value	0.0497	0.5659	0.8815	0.6314	0.4921	0.6936	0.7517
41-77	Algo	5.6381	7.292	7.7138	6.0926	7.1689	7.6914	7.1892
	Random	6.0963	8.608	8.3684	7.2037	7.9128	8.6171	8.3181
	p-value	0.3728	0.2038	0.2871	0.4044	0.2313	0.4268	0.1512

We observe that the age interval with the highest numbers of semi matches is 24-29, while the oldest individuals have the lowest instances. Moreover, for both females and males the most unpopular cluster (table 2 and 4) is the first (introvert and disorganized). The females in cluster 2 (extrovert and non-athlete) and the males who fall under "partying geek" and "social organized" have the highest number of semi matches.

Table 4: Mean semi matches for each male cluster

		Clusters						
Age		**1**	**2**	**3**	**4**	**5**	**6**	**7**
18-23	Algo	0.1860	0.1993	0.2030	0.1961	0.2230	0.2058	0.1875
	Random	0.1799	0.1963	0.1935	0.2162	0.1934	0.2078	0.1897
	p-value	0.7517	0.8266	0.5995	0.1736	0.0629	0.8933	0.9141
24-29	Algo	0.1966	0.2462	0.2250	0.2513	0.2490	0.2370	0.2492
	Random	0.1977	0.2332	0.2166	0.2263	0.2419	0.2091	0.2346
	p-value	0.9376	0.2108	0.5139	0.0216	0.4869	0.0103	0.2875
30-40	Algo	0.2008	0.2106	0.1954	0.2346	0.2281	0.2346	0.2256
	Random	0.1853	0.2122	0.2054	0.2404	0.2115	0.2107	0.2021
	p-value	0.2091	0.8928	0.3590	0.6777	0.0698	0.0106	0.0554
41-77	Algo	0.1468	0.1574	0.1567	0.1472	0.1812	0.1736	0.1650
	Random	0.1459	0.1668	0.1591	0.1353	0.1525	0.1758	0.1587
	p-value	0.9591	0.6553	0.8427	0.6934	0.0295	0.9327	0.6949

3 Developing a formal model

3.1 The First Step: The Bias (from exposure to semi match)

Following a review of the seeker groups and the changes experienced by them under the algorithm, we explore the role exposure plays in facilitating semi matches more generally. Understanding the role exposure plays serves as an indication of how much the algorithm relies on the quantity of suggestions as opposed to their quality (see figure 2 again for channels).

1. In the first model (column 1 of table 5) we begin by modelling the relationship between a seeker's exposure and their semi matches in both random and algorithmic cases. Additional explanatory variables are added gradually to detect whether the exposure coefficient changes.

2. In the second column the model includes the seeker's gender, their personality cluster and the age group. The third column adds their interactions.

3. The fourth version adds covariates detailing the type of recommendation each seeker had on average. This is measured by gendermatch (the percentage of recommendations in which a seeker was suggested to the same gender) and agematch (same principle applied to age groups).

4. In the fifth and final specification we include social proximity measures, six variables that note the absolute mean distance a seeker has to their listers across all recommendations on each of the six scales. To illustrate, consider a seeker with a score of seven on the "organised" scale and a lister with a score of four; the proximity measure

generated is three. In the model we use the averages of each seeker.

The model specification is as follows:

$$
\begin{aligned}
\#\text{Semi matches} = \beta_0 + \beta_1\text{Exposure} &+ \beta_2\text{Seeker Controls} \\
&+ \beta_3\text{Recommendation Controls} \\
&+ \beta_4\text{Social Proximity Measure} + \epsilon
\end{aligned}
$$

Table 5 shows that on average the random engine generates more semi matches than the algorithmic one. The interaction of the random variable with exposure illustrates the fact that the same amount of suggestions leads to fewer semi matches under the algorithm. Not only is the algorithm less generous in suggesting seekers, it is also suggesting them to listers that are less interested in them. A further noteworthy detail is the fact that female seekers exhibit an increased probability of getting a semi match as opposed to their male counterparts. In addition, the interaction of gender with the random engine, while positive, is not significant - indicating again that women are not subjected to a bias.

The coefficients for the age groups (not specified above) while significant, state that the algorithm does a better job of attaining semi matches than the random engine. Once more, we observe no systematic discrimination on the grounds of either gender or age groups. The model confirms the results in the tables.

Table 5: Semi match and Exposure

	(1)	(2)	(3)	(4)	(5)
	Semi match	Semi match	Semi match	Semi match	Semi match
Exposure	0.261***	0.265***	0.265***	0.265***	0.265***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Random	0.144*	0.109*	0.162**	0.224***	0.153*
	(0.041)	(0.038)	(0.031)	(0.024)	(0.049)
Random*Exposure	-0.029**	-0.031*	-0.031**	-0.031**	-0.031**
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Gender (Female=1)		0.821***	0.848***	0.853***	0.854***
		(0.023)	(0.001)	(0.002)	(0.003)
Random×Gender		0.112*	-0.001	0.007*	0.010
		(0.031)	(0.002)	(0.003)	(0.005)
Clusters		yes	yes	yes	yes
Random×Cluster		yes	yes	yes	yes
Age group		yes	yes	yes	yes
Random×Age group		yes	yes	yes	yes
Cluster×Age Group (28)			yes	yes	yes
R×C×A			yes	yes	yes
Cluster×Gender (14)			yes	yes	yes
R×C×G			yes	yes	yes
Age Group×Gender(8)			yes	yes	yes
R×A×G			yes	yes	yes
Cluster×Age×Gender (56)			yes	yes	yes
R×C×A×G			yes	yes	yes
Gender Match				-0.042$^{(*)}$	-0.043*
				(0.014)	(0.016)
R×Gender Match				-0.063*	-0.066*
				(0.018)	(0.019)
Lister Gender				-0.124***	-0.131***
				(0.018)	(0.016)
R×Lister Gender				-0.081**	-0.085**
				(0.018)	(0.019)
Age Match				yes	yes
R×Age Match				yes	yes
Proximity(6)					yes
R×Proximity(6)					yes
Constant	0.137***	-0.425***	-0.399***	-0.318***	-0.193*
	(0.017)	(0.045)	(0.028)	(0.027)	(0.064)
R-squared	52.6%	55.95%	56.02%	56.08%	56.17%
Observations	87,494	87,494	87,494	87,494	86,801

R - Random; C - Cluster; A - Age group; G - Gender

Standard Errors calculated at cluster level

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.2 The Second Step: The Efficiency (from semi match to final match)

The model for the second stage brings into focus the relationship between semi matches and the number of final matches a seeker obtains. Following the style employed in the first stage, the relationship between the final matches and semi matches is accompanied by the same covariates as those from table 6.

The model specification for the second step:

$$\begin{aligned}\#\text{Final matches} = \beta_0 + \beta_1\#\text{Semi matches} &+ \beta_2\text{Seeker Controls}\\ &+ \beta_3\text{Recommendation Controls}\\ &+ \beta_4\text{Social Proximity Measure} + \epsilon\end{aligned}$$

Table 6 indicates that for every additional semi match a seeker obtains, their chances of receiving a final match are increased. A positive coefficient is unsurprising as semi matches are required for final matches to form. Nevertheless, the question of whether the importance of semi matches changes under the algorithm is pertinent one, as it indicates the channel the algorithm relies on (quality vs. quantity). The interaction between the random dummy and the semi match variable is non-significant, suggesting that the algorithm does not rely on the quality channel.

Another remarkable trend is that for females, the algorithm is less effective in increasing the final matches through the number of semi matches. One of the potential explanations could be that, as seen in table 5, females have an overall higher chance of getting semi matches under the algorithm. As a result, they have a larger pool of listers to choose from and may therefore be more selective. The relationship between the number of interested users a person has and rigour of their selections is a topic that merits attention. In table 6 we only included the final model.

Table 6: Final Match and semi match

	(1) Final Match
Semi match	0.167***
	(0.007)
Random	-0.019
	(0.010)
Random×Semi match	-0.003
	(0.004)
Gender (Female=1)	-0.093***
	(0.005)
Random×Gender	0.049***
	(0.004)
Clusters	yes
Random×Cluster	yes
Age group	yes
Random×Age group	yes
Cluster×Age Group (28)	yes
R×C×A	yes
Cluster×Gender (14)	yes
R×C×G	yes
Age Group×Gender(8)	yes
R×A×G	yes
Cluster×Age×Gender (56)	yes
R×C×A×G	yes
Gender Match	-0.010
	(0.004)
R×Gender Match	0.028**
	(0.007)
Lister Gender	0.028*
	(0.008)
R×Lister Gender	0.017
	(0.009)
Age Match	yes
R×Age Match	yes
Proximity(6)	yes
R×Proximity(6)	yes
Constant	-0.007
	(0.009)
R-squared	26.4%
Observations	43747

R - Random; C - Cluster; A - Age group; G - Gender

Standard Errors calculated at cluster level

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4 Distributional Considerations

4.1 The Utility of Market Access

Up until now, the paper has attempted to determine the quality of the service offered by Badi as a function of the semi matches or final matches. While this gives a broad overview of how the algorithm performs, a set of normative aspects commonly discussed in the *context of market access* merit closer attention. The notion revolves around the utility a seeker experiences from a given match. Would it be realistic to infer that the utility of the first final match is akin to that of the fifth one? Or the 15th? This section reviews the outcomes of the two recommender engines from a utilitarian and distributional standpoint.

Resting on the assumption that a service leading to a higher number of final matches is superior, we introduce the utility function:

$$Utility = 3 \times (x)^{0.25}$$

where x is the number of final matches or market access (for the sake of brevity, we do not consider prices or amenities or whether they finally rent a room). We observe the following properties:

1. A seeker's utility in using the matching service is independent of the outcome of other seekers or listers and is a direct result of the number of final matches he or she experiences.
2. The function has outputs in the range $[0, +\infty)$ and is continuously differentiable
3. Utility is monotonically increasing, with U(0) = 0.
4. The limit of the first derivative is positive infinity as x approaches zero, and zero as the input tends to infinity, complying with Inada conditions (Inada, 1963)

In attributing a utility to each seeker in the market, we consider the overall welfare of the users on the platform under the random and algorithmic engines. Assuming the utilitarian stance that each individual is equally important:

$$Welfare = \sum_{i=1}^{N} Utility_i$$

We consider how efficiently resources were allocated in maximising the level of welfare. Table 7 shows that the average number of final matches as well as the mean utility experienced by a seeker in the algorithmic case is lower.

Given that the marginal utility is decreasing, we calculate the amount of final matches that were "misallocated" in contributing to the current level of welfare. While the recommendation engine only has partial control over the number of final matches, the essence of the argument rests on the fact that the algorithm is further exacerbating differences in utility, leading to suboptimal levels of welfare.

Table 7: Welfare function

	Engine	
Welfare	Random	Algo
E(Final Matches)	0.2725	0.2602
Utility(E(Final Matches))	2.1676	2.1427
E(Utility)	0.5250	0.4967
EDEM	0.0009	0.0007
Loss in matches	0.2716	0.2595

The equally distributed equivalent number of matches (EDEM) is inspired by the EDEI measure used in the literature on income inequality (Atkinson, 1970). EDEM is the number of matches that every individual in a group would have to receive to achieve the same degree of welfare that is observed in the existing unequal distribution. A larger EDEM indicates a more equal society and less wasted resources. As is visible in the table above, the algorithm is less equitable in terms of market access.

4.2 The distributional analysis (Concentration curve)

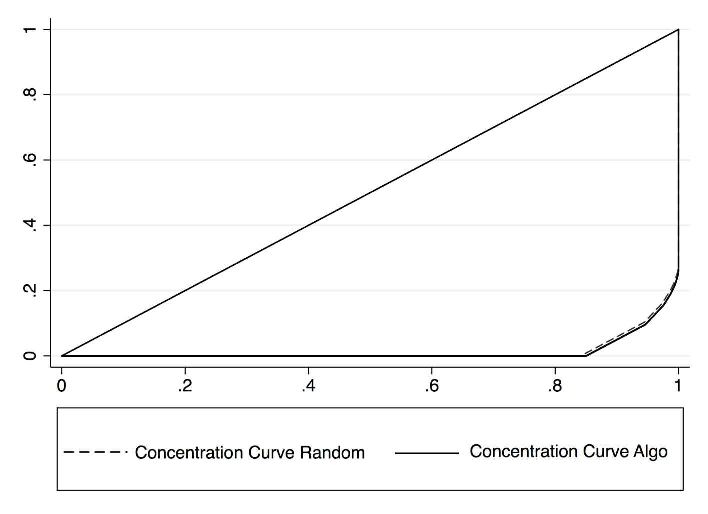

Figure 6: Concentration curves

There are a number of shortcomings with measuring inequality through the utilitarian approach. It is limited in solely depending on the final matches a person receives and rests on the rather strong assumption that every individual has the same utility function (Sen, 1979). We respond to this shortcoming by exploring the level of inequality by means of a concentration curve (Lorenz, 1905). Similar to the Lorenz curve, this indicator maps the cumulative distribution of individuals to their cumulative final matches. The graph orders individuals by their “wealth” or number of matches, meaning the poor are counted first. A curve close to the 45 line indicates that wealth increases at the same rate at which individuals are counted, implying an equal society. The stronger the deviation from the 45° line, the larger the gap between the rich and the poor. As depicted in Figure 6 above, the concentration curve for the algorithm deviates more from the line than the random one, confirming the inequality concerns first raised by the EDEM measure. For both engines, we observe that a staggering 85% of the population experiences no final matches.

5 Discussion and Conclusion

Within the Economics literature, discrimination has mostly been considered a product of unethical individual preferences and has focused on policy implementations that mitigate the disadvantages faced by subgroups (Neumark et al., 1996; Bertrand and Mullainathan, 2004; Goldin and Rouse, 2000) particularly in the area of labour market policies (Deakin et al., 2005; Siegelman and Heckman, 1993; Turner et al., 1991; Becker, 1971). The prospect of such harmful trends being augmented through algorithmic optimization processes is a notion that has mostly captured the attention of researchers in the fields of computer science and machine learning (Crawford and Schultz, 2014; Crawford et al., 2014; Kirkpatrick, 2016; Hajian et al., 2016; Richterich, 2018), its implications not having been considered from an economic standpoint. Given the increased use and dependence of markets on algorithmic optimisation processes (Rauch and Schleicher, 2015) this paper makes a first attempt to introduce the notion of bias and outline the economic conditions in which it appears, as well as enquiring into a potential efficiency-equity tradeoff.

To our surprise, the algorithm underperformed the random recommendation engine in almost every measure it was subjected to. With the exception of a few arbitrary groups, it generally exposed seekers less - yet proved relatively ineffective in using the alternative channel of quality to increase a seeker’s semi matches. The second stage model served to show that the semi matches generated by the algorithm were in no way superior to those resulting from the random engine, as would be the case if the final match count went up. Within the confines of the dataset provided, we find no reason to believe that the algorithm is offering an overall superior service.

Nevertheless, it is worth mentioning that our models rely on a number of assumptions that were made due to the constraints imposed by the data. As the information used for this analysis is a fraction of the entire activity present on Badi's network, it is impossible to estimate the bounds of the number of suggestions that can be answered by listers. The zero sum restriction is therefore an existing but vague limitation that hinders us from making a statement on whether *all* resources were used (i.e. were all the suggestions that could be answered, sent out?).

Additionally, the data used in this analysis does not include all activity related to a particular user. If a seeker is suggested above average it is considered that the algorithm does increase his exposure, however in the absence of all recommendations being available it is unfeasible to state how representative the increase truly is. We assume that all recommendations had the same likelihood of being selected for our data, and therefore our conclusions do mirror the larger trends. This restriction can only be overcome by using all available data- a decision that ultimately rests with Badi.

The suggestion and matching process involves one more aspect that proved challenging to quantify and was therefore omitted. The majority of profiles contain a picture of the user, and given the possibility of physical appearance or other features affecting the decision of accepting someone as a potential roommate (Pope and Sydnor, 2011) - this could lead to an omitted variable bias in our model if the picture features correlate with the other variables. While we consider this an unlikely scenario, it could be a reason behind the models' reduced explanatory powers (our R^2 was around 56% for the first stage and 25% for the second).

In spite of these limitations, our approach and conclusions remain consequential: systematic discrimination continues to gain importance and the measures used in identifying it in markets play an important first step in tackling its effects. A topic that requires a comprehensive debate is the extent to which an algorithm is responsible for minimizing existing discrimination, and the viability of such a task. How much can a matching mechanism correct? What laws must be implemented to ensure an equitable access to the market while still improving the utility of each individual? As the algorithm reviewed here presented no efficiency gains, this question could not be answered fully. The methods presented above may provide a foundation for the development of ethical algorithms in the context of matching markets.

References

Albinsson, P. A. and Yasanthi Perera, B. (2012). Alternative marketplaces in the 21st century: Building community through sharing events. *Journal of consumer Behaviour*, 11(4):303–315.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 23.

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*, 2(3):244–263.

Becker, G. (1971). The economics of discrimination 2nd ed (university of chicago press, chicago).

Bertrand, M. and Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013.

Crawford, K. (2016). Artificial intelligence's white guy problem. *The New York Times*.

Crawford, K., Gray, M. L., and Miltner, K. (2014). Big data— critiquing big data: Politics, ethics, epistemology— special section introduction. *International Journal of Communication*, 8.

Crawford, K. and Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.*, 55:93.

Deakin, S. F., Morris, G. S., and Morris, G. S. (2005). *Labour law*. Hart Publishing Oxford.

Debreu, G. and Scarf, H. (1963). A limit theorem on the core of an economy. *International Economic Review*, 4(3):235–246.

Dutta, B. and Massó, J. (1997). Stability of matchings when individuals have preferences over colleagues. *Journal of Economic Theory*, 75(2):464–475.

Einav, L., Farronato, C., and Levin, J. (2016). Peer-to-peer markets. *Annual Review of Economics*, 8:615–635.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417.

Fradkin, A. (2017). Search, matching, and the role of digital marketplace design in enabling trade: Evidence from airbnb.

Gale, D. and Shapley, L. S. (1962). College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15.

Goldin, C. and Rouse, C. (2000). Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American Economic Review*, 90(4):715–741.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F. Å., and Hansen, L. K. (1999). On clustering fmri time series. *NeuroImage*, 9(3):298–310.

Hajian, S., Bonchi, F., and Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2125–2126.

Hamari, J., Sjöklint, M., and Ukkonen, A. (2016). The sharing economy: Why people participate in collaborative consumption. *Journal of the Association for Information Science and Technology*, 67(9):2047–2059.

Inada, K.-I. (1963). On a two-sector model of economic growth: Comments and a generalization. *The Review of Economic Studies*, 30(2):119–127.

Kirkpatrick, K. (2016). Battling algorithmic bias: how do we ensure algorithms treat us fairly? *Communications of the ACM*, 59(10):16–17.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American statistical association*, 9(70):209–219.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Makles, A. (2012). Stata tip 110: How to get the optimal k-means cluster solution. 12.

Malkiel, B. G. (2003). The efficient market hypothesis and its critics. *Journal of economic perspectives*, 17(1):59–82.

Neumark, D., Bank, R. J., and Van Nort, K. D. (1996). Sex discrimination in restaurant hiring: An audit study. *The Quarterly journal of economics*, 111(3):915–941.

Niederle, M., Roth, A. E., and Sönmez, T. (2008). Matching. *The New Palgrave Dictionary of Economics. Palgrave Macmillan*, 2.

Pope, D. G. and Sydnor, J. R. (2011). What's in a picture? evidence of discrimination from prosper. com. *Journal of Human Resources*, 46(1):53–92.

Rauch, D. E. and Schleicher, D. (2015). Like uber, but for local government law: The future of local regulation of the sharing economy. *Ohio St. LJ*, 76:901.

Richterich, A. (2018). The big data agenda: Data ethics and critical data studies.

Roth, A. E. (1982). The economics of matching: Stability and incentives. *Mathematics of operations research*, 7(4):617–628.

Roth, A. E., Sönmez, T., and Ünver, M. U. (2004). Kidney exchange. *The Quarterly Journal of Economics*, 119(2):457–488.

Roth, A. E. and Sotomayor, M. (1992). Two-sided matching. *Handbook of game theory with economic applications*, 1:485–541.

Sen, A. (1979). Utilitarianism and welfarism. *The Journal of Philosophy*, 76(9):463–489.

Shapley, L. and Scarf, H. (1974). On cores and indivisibility. *Journal of mathematical economics*, 1(1):23–37.

Siegelman, P. and Heckman, J. (1993). The urban institute audit studies: Their methods and findings. *Clear and Convincing Evidence: Measurement of Discrimination in America, Washington*, 187:258.

Sweeney, L. (2013). Discrimination in online ad delivery. *Queue*, 11(3):10.

Turner, M. A., Fix, M., and Struyk, R. J. (1991). *Opportunities denied, opportunities diminished: Racial discrimination in hiring.* The Urban Insitute.

Appendix

There are some priors defined in the random engine that changes the probability of an individual being recommended by the random engine. To account for this endogeneity, we first estimate a differenced model where the dependent variable is the change in the number of semi matches from algorithm to random and the variable of interest is the change in the number of times an individual is exposed from algorithm to random. We run multiple specifications with various controls to see if the coefficient is robust across models.

Table 8: Semi matches and Exposure

	(1)	(2)	(3)	(4)	(5)
	ΔSemi match	ΔSemi match	ΔSemi match	ΔSemi match	ΔSemi match
ΔExposure	0.253^{***}	0.253^{***}	0.253^{***}	0.253^{***}	0.253^{***}
	(25.26)	(25.51)	(25.46)	(25.50)	(25.20)
Gender (Female=1)		-0.131^{**}	-0.217^{***}	-0.222^{***}	-0.217^{***}
		(-5.08)	(-112.98)	(-94.95)	(-141.2)
Clusters		yes	yes	yes	yes
Age group		yes	yes	yes	yes
C × A (28)			yes	yes	yes
C × G (14)			yes	yes	yes
A × G (8)			yes	yes	yes
C × A × G (56)			yes	yes	yes
ΔGender Match				yes	yes
ΔLister Gender				yes	yes
ΔAge Match				yes	yes
ΔProximity(6)					yes
Constant	0.0334	0.0377	-0.0188^{***}	-0.0166^{***}	-0.0165^{***}
	(1.75)	(1.81)	(-13.08)	(-13.14)	(-14.08)
R-squared	32.90%	33.03%	33.13%	3319%	33.23%
Observations	43747	43747	43747	43747	43058

C - Cluster; A - Age group; G - Gender

Standard Errors calculated at cluster level

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The above table shows that adding multiple controls does not alter the coefficient. Thus, we can infer that exposure in itself has a very important role and does not change with the type of controls that are included in the model. The differenced model also provides an estimate of how an additional exposure has a different impact on the semi matches based on the engine by which the recommendation is generated, the gender of the seeker, the age group and the cluster.

The below table provides us the detailed description of the ”types” of individuals in each cluster.

Table 9: Cluster composition

Cluster	Partier	Athlete	Organiser	Active	Geek	Social
Cluster1	0.12	0.34	0.89	0.35	0.09	0.67
Cluster2	5.44	4.99	7.86	7.78	0.94	8.46
Cluster3	1.15	1.51	8.00	5.41	0.47	7.41
Cluster4	5.14	7.12	7.68	8.14	5.84	8.62
Cluster5	1.29	7.84	8.73	8.17	0.82	8.59
Cluster6	2.08	3.41	6.69	5.54	6.34	6.78
Cluster7	1.60	6.23	4.03	5.28	0.90	6.28

The below graph shows how we can use k-means to get the desired number of clusters by the elbow method.

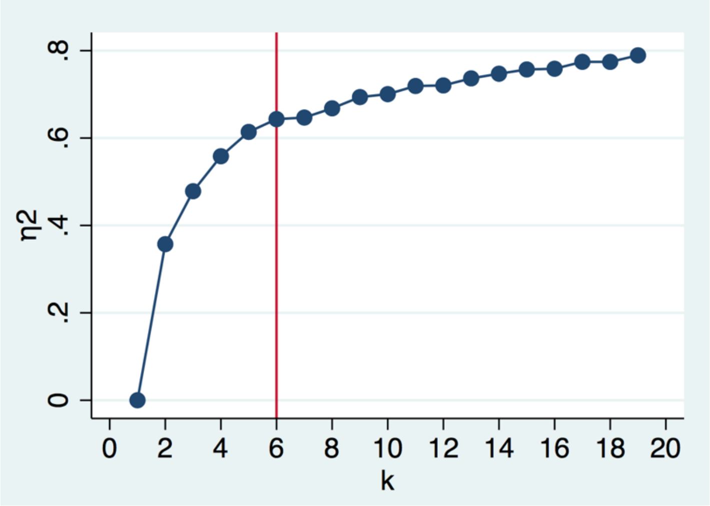

Figure 7: K-means and elbow method