arXiv:1902.00375v1 [cs.LG] 1 Feb 2019

Dynamic fairness – Breaking vicious cycles in automatic decision making

Benjamin Paaßen, Astrid Bunge, Carolin Hainke, Leon Sindelar, and Matthias Vogelsang

CITEC Center of Excellence, Bielefeld University*

Preprint of the ESANN 2019 paper Paaßen et al. [2019] as provided by the authors. The original can be found at the ESANN electronics proceedings page.

Abstract

In recent years, machine learning techniques have been increasingly applied in sensitive decision making processes, raising fairness concerns. Past research has shown that machine learning may reproduce and even exacerbate human bias due to biased training data or flawed model assumptions, and thus may lead to discriminatory actions. To counteract such biased models, researchers have proposed multiple mathematical definitions of fairness according to which classifiers can be optimized. However, it has also been shown that the outcomes generated by some fairness notions may be unsatisfactory.

In this contribution, we add to this research by considering decision making processes in time. We establish a theoretic model in which even perfectly accurate classifiers which adhere to almost all common fairness definitions lead to stable long-term inequalities due to vicious cycles. Only demographic parity, which enforces equal rates of positive decisions across groups, avoids these effects and establishes a virtuous cycle, which leads to perfectly accurate and fair classification in the long term.

Automatic decision-making via machine learning classifiers carries the promise of quicker, more accurate, and more objective decisions because automatic mechanisms do not foster animosity against any group [Munoz et al., 2016, O'Neil, 2016]. Yet, machine learning systems can indeed reproduce and exacerbate bias that is encoded in the training data or in flawed model assumptions [Munoz et al., 2016, O'Neil, 2016, Corbett-Davies and Goel, 2018, Dwork et al., 2012]. For example, the COMPAS tool, which estimates the risk of recidivism of defendants in the US law system prior to trial, has been found to have higher rates of false positives for Black people compared to white people and has thus been called unfair [Angwin et al., 2016]. Similarly, a tool developed by Amazon to rate the résumés of job applicants assigned higher scores to men compared to women because successful applicants in the past had mostly been male [Dastin, 2018]. Finally, multiple machine-learning-based credit scoring systems have emerged that reproduce historical biases and systematically assign lower credit scores to members of disenfranchised minorities [O'Dwyer, 2018].

In general, we consider scenarios where individuals $i \in \{1, \ldots, m\}$ in some population of size m apply for some positive outcome, such as a pre-trial bail, a job, or a loan, and a gatekeeper institution decides whether to grant that outcome, with the interest of accepting only those individuals who will "succeed" with that outcome, e.g. not commit a crime, succeed

*Funding by the CITEC center of excellence (EXC 277) is gratefully acknowledged.

in their job for the company, or pay back a loan. To make that decision, the institution employs a binary classifier $f : \{1, \ldots, m\} \to \{0, 1\}$ that predicts whether to grant the outcome, i.e. $f(i) = 1$, or not, i.e. $f(i) = 0$. Now, let $y_i \in \{0, 1\}$ denote whether an individual will succeed ($y_i = 1$) or not ($y_i = 0$). Then, the aim of the classifier is to maximize the share of the population where $f(i) = y_i$.

In our examples, we care how a certain *protected group* C is treated compared to everyone else. In general, we assume these protected groups to be pre-defined by society, e.g. via the EU charter of fundamental rights, which forbids discrimination based on sex, race, color, ethnic or social origin, religion, political opinion, and several other features [for Fundamental Rights, 2012]. Formally, let $C \subseteq \{1, \ldots, m\}$ be a protected group, let $m_c := |C|$, let $\neg C := \{1, \ldots, m\} \setminus C$, and let $m_{\neg c} := |\neg C|$. Then, the fairness notion corresponding to our last two examples is *demographic parity*, which requires that the rate of positive decisions is equal across groups, i.e. $\sum_{i \in C} \frac{f(i)}{m_c} = \sum_{i \in \neg C} \frac{f(i)}{m_{\neg c}}$ [Dwork et al., 2012, Žliobaitė, 2017].

Multiple authors have criticized demographic parity because it decreases accuracy if the base rate of successful people is different across groups [Corbett-Davies and Goel, 2018, Žliobaitė, 2017, Hardt et al., 2016]. Accordingly, Hardt et al. have proposed the notion of *equalized odds* which only requires an equal rate of positive decisions among the people who will succeed and the people who will not succeed [Hardt et al., 2016], which corresponds to the fairness notion in the COMPAS example [Angwin et al., 2016].

In addition distributive justice considerations, several authors have proposed notions of due process, in the sense that any classifier should be considered fair which performs decisions in a fair way [Grgić-Hlača et al., 2016]. In particular, several authors have argued that classifiers should not use features that code the protected group directly or indirectly [O'Neil, 2016, Dwork et al., 2012, Grgić-Hlača et al., 2016, Kilbertus et al., 2017, Kusner et al., 2017]. Alternatively, Corbett and Goel have proposed a two-step classification process. First, a function $g : \{1, \ldots, m\} \to \mathbb{R}$ assigns a risk score to each individual, which should increase monotonously with the probability to be successful, i.e. $g(i) = \sigma(P(y_i = 1))$ for some monotonous function σ (a property also called *calibration* [Liu et al., 2018]). Second, the actual classifier should be a thresholding on the risk score, i.e. $f(i) = 1$ if $g(i) \geq \theta$ for some fixed threshold $\theta \in \mathbb{R}$ and $f(i) = 0$ otherwise, thus holding everyone to the same standard [Corbett-Davies and Goel, 2018].

In this contribution, we argue that even if a classifier is perfectly accurate and is fair according to all fairness notions except demographic parity, we may still obtain undesirable long-term outcomes. To do so, we establish a simple dynamical system which assumes that positive classifier decisions have positive impact on the future success rate of a group, which in turn leads to a higher chance for positive classifications and so on. We show that this positive feedback loop implies stable equilibria where a protected group receives no positive decisions anymore. We also show that imposing demographic parity breaks this feedback loop and introduces a single, stable equilibrium which exhibits perfect accuracy, equality, and fairness according to all notions.

Our model is inspired by prior work of O'Neil, who has investigated existing automatic decision making systems and found positive feedback loops which disadvantage protected groups [O'Neil, 2016]. However, O'Neil did not provide a theoretic model. Further, our work is related to prior research by Liu et al., who have analyzed one-step dynamics in a credit scoring scenario [Liu et al., 2018] but did not consider long-term outcomes. Finally, Hu and Chen have previously analyzed a detailed economic model of the labor market, including long-term dynamics [Hu and Chen, 2018] and found that demographic parity lead to a desirable equilibrium. Our model can be seen as complementary to theirs, as we focus more generally on settings where classifier decisions impact objective risk, whereas their investigation was focuses specifically on the labor market. We also provide a stronger analysis in the sense that our model predicts undesirable equilibria even in case the classifier is perfectly accurate and fair to multiple established criteria. As such, we provide, to our knowledge, the first theoretic model which explicitly covers long-term dynamics of automatic decision making for perfectly accurate and fair classifiers.

1 Model

In our model, we assume that every individual i has an objective risk score q_i^t at time t which is drawn from an exponential distribution[1] with mean μ_c^t if $i \in C$ and with mean $\mu_{\neg c}^t$ otherwise. Further, we assume that the $n \leq m$ people with the highest score in each iteration are the ones which will be successful, i.e. $y_i = 1$ if and only if q_i^t is among the top n at time t. Accordingly, we obtain a perfectly accurate classifier if we use the scoring function $g^t(i) = q_i^t$ and set the decision threshold θ^t such that exactly the top n scores are above or equal to it. Note that our hypothetical classifier conforms to equalized odds because there are no misclassifications [Hardt et al., 2016], fulfills the calibration, threshold, and accuracy requirements of Corbett and Goel [Corbett-Davies and Goel, 2018], and does not need access to the group label, neither directly nor indirectly, thus conforming to all due process notions of fairness [O'Neil, 2016, Dwork et al., 2012, Grgić-Hlača et al., 2016, Kilbertus et al., 2017, Kusner et al., 2017].

We estimate the overall number of people who receive a positive classification decision inside and outside the protected group via the expected values $\mathbb{E}[\sum_{i \in C} f(i)] = m_c \cdot \int_{\theta^t}^{\infty} \frac{1}{\mu_c^t} \cdot \exp(-\frac{q}{\mu_c^t}) dq = m_c \cdot \exp(-\frac{\theta^t}{\mu_c^t})$ and $\mathbb{E}[\sum_{i \in \neg C} f(i)] = m_{\neg c} \cdot \exp(-\frac{\theta^t}{\mu_{\neg c}^t})$[2].

We finally assume that the mean for a group improves with a higher rate of positive classifier decisions in the previous time step according to the following equation.

$$\begin{pmatrix} \mu_c^{t+1} \\ \mu_{\neg c}^{t+1} \end{pmatrix} = (1 - \alpha) \cdot \begin{pmatrix} \mu_c^t \\ \mu_{\neg c}^t \end{pmatrix} + \beta \cdot \begin{pmatrix} \exp(-\frac{\theta^t}{\mu_c^t}) \\ \exp(-\frac{\theta^t}{\mu_{\neg c}^t}) \end{pmatrix} \tag{1}$$

where the decision threshold θ^t is selected as the numeric solution to the equation $n = m_c \cdot \exp(-\frac{\theta^t}{\mu_c^t}) + m_{\neg c} \cdot \exp(-\frac{\theta^t}{\mu_{\neg c}^t})$, where $\alpha \in [0, 1]$ is a hyper-parameter describing how much score an individual loses in each time step ("leak reate"), and where $\beta \in \mathbb{R}^+$ is a hyper-parameter describing how much score an individual gains for for each positive classifier decision. Figure 1 (left) visualizes the dynamical system.

Note the connections of our model to the real-world examples mentioned before. In credit scoring, q_i^t would correspond to the credit score, i.e. the capability of an individual to pay back a loan. We would plausibly assume that the score increases with positive classifier decisions because individuals who get a loan have additional financial resources at their disposal and can use those to add wealth to their group [Liu et al., 2018]. Further, we would assume a nonzero leak rate α because individuals need to cover their expenses which may negatively affect their capability to pay back a loan.

[1] Note that our qualitative results can be generalized to other distributions, such as Gaussian or Pareto. We select the exponential distribution here because it only has a single parameter and thus is easier to analyze. You can find the full analysis in Appendix A.

[2] We consider each classifier decision as a Bernoulli trial with success probability $P = \int_{\theta^t}^{\infty} \frac{1}{\mu_c^t} \cdot \exp(-\frac{q}{\mu_c^t}) dq$, yielding a binomially distributed random variable $\sum_{i \in C} f(i)$ with expected value $m_c \cdot P$ and variance $m_c \cdot P \cdot (1 - P)$. Note that the variance gets close to zero if P is small itself, such that the expected value is a precise estimate for sufficiently small n.

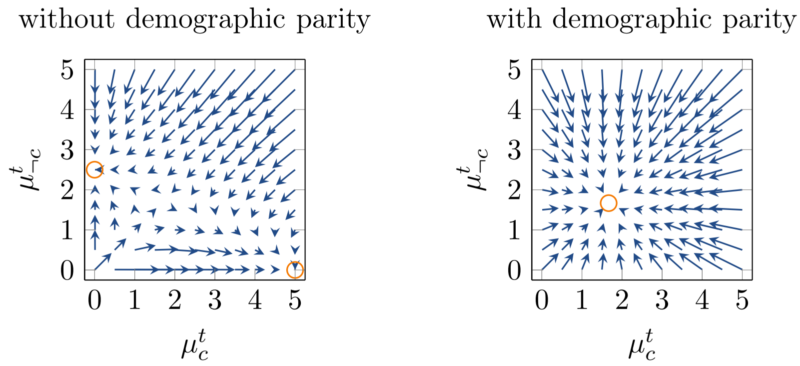

Figure 1: An illustration of the dynamical system model from Equation 1 for a population with $m_c = 100$, $m_{\neg c} = 200$, $n = 50$ successful people, leak rate $\alpha = 0.5$, and score $\beta = 5$. Equilibria are highlighted with circles. Left: The model without demographic parity requirement, exhibiting undesirable stable equilibria at the coordinate axes. Right: The model with demographic parity, exhibiting a single stable equilibrium on the diagonal.

If we apply our model to pre-trial bail assessment, the score q_i^t would assess the likelihood of a defendant to not commit a crime until trial. Here, we would assume that the score decreases with negative classifier decisions because incarcerating people from a community may cut social ties and deteriorate trust in the state, leading to a higher crime rate [O'Neil, 2016]. This effect can be modeled by a nonzero leak rate α and a positive score β.

Also note that our model is not necessarily realistic but shows that there exist contexts where even perfect classifiers can exhibit stable long-term inequality. We show that context can matter, not that every context conforms to our model.

If we analyze the equilibria of this system, we first note that $\lim_{\mu_c^t \to 0} \mu_c^{t+1} = \lim_{\mu_c^t \to 0} (1 - \alpha) \cdot \mu_c^t + \beta \cdot \exp(-\frac{\theta^t}{\mu_c^t}) = 0$, i.e. $\mu_c^* = 0$ is a fix point. Further, $\lim_{\mu_c^* \to 0} \exp(-\frac{\theta^*}{\mu_c^*}) = 0$, i.e. no person from the protected group is above the threshold at that fix point. Accordingly, we can compute the fix point threshold θ^* only for the non-protected group, i.e. $\theta^* = \mu_{\neg c}^* \cdot \log(\frac{m_{\neg c}}{n})$. By plugging this into the fix point equation $\mu_{\neg c}^* = (1 - \alpha) \cdot \mu_{\neg c}^* + \beta \cdot \exp(-\frac{\theta^*}{\mu_{\neg c}^*})$ we obtain $\mu_{\neg c}^* = \frac{\beta}{\alpha} \cdot \frac{n}{m_{\neg c}}$, which yields $\mu_{\neg c}^* = 2.5$ for our example in Figure 1 (left). At this fix point, we obtain a Jacobian of Equation 1 which is $1 - \alpha$ times the identity matrix, i.e. both eigenvalues have an absolute value < 1 for $\alpha > 0$, implying stability. In Figure 1 (left) we also see that the basin of attraction is the entire region above the diagonal, i.e. whenever we start with slight inequality in favor of the non-protected group, this inequality will get amplified.

In summary, we have shown that, for our exponential distribution model, there are always undesirable and stable equilibria in which μ_c^t degenerates to zero and the non-protected group receives all positive outcomes. This begs the question: Can we break this undesirable dynamic? Indeed, we can, using demographic parity.

2 Demographic Parity Dynamics

Demographic parity requires equal acceptance rates across groups, i.e. $\exp(-\frac{\theta_c^t}{\mu_c^t}) = \exp(-\frac{\theta_{\neg c}^t}{\mu_{\neg c}^t}) = P$ for some acceptance rate P and group-specific thresholds θ_c^t and $\theta_{\neg c}^t$. We obtain P as solution of the threshold equation $n = m_c \cdot P + m_{\neg c} \cdot P$, i.e. $P = \frac{n}{m_c + m_{\neg c}} = \frac{n}{m}$. By plugging this result into our fix point equation we obtain $\mu^* = \mu_c^* = \mu_{\neg c}^* = (1 - \alpha) \cdot \mu^* + \beta \cdot P = \frac{\beta}{\alpha} \cdot \frac{n}{m}$, which yields $\mu^* = 5/3$ for our example in Figure 1 (right). For this fix point we obtain a Jacobian of Equation 1 of $1 - \alpha$ times the identity matrix, implying stability.

Overall, demographic parity ensures that the mean for every group converges to the same point, such that the thresholds θ_c^t and $\theta_{\neg c}^t$ become equal as well. This, in turn, implies that selecting the top-scored people in each group corresponds to selecting the top-scored people in the entire population, implying a classifier that is perfectly accurate *and* conforms to all notions of fairness, including demographic parity.

3 Conclusion

In this contribution, we have analyzed a simple dynamic model for automatic decision making. In particular, our model assumes that people should receive a positive classifier decision only if their objective risk score is in the top, that the means of the score distribution differ between the protected group and everyone else, and that positive decisions improve the mean for the group in the next time step. This feedback loop becomes a vicious cycle in which even a perfectly accurate classifier conforming to almost all fairness notions leads to stable inequality. Fortunately, we can break this vicious cycle by imposing democratic parity which instead leads to an equilibrium with perfectly accurate, equal, and fair classification.

At present, our analysis is limited to a theoretical model assuming an exponential distribution and a simple dynamic model. However, we note that generalizations to other distributions are possible. Further, we note that our findings are consistent with practical application scenarios [O'Neil, 2016] and more realistic as well as detailed models for the labor market [Hu and Chen, 2018].

Overall, we conclude that our findings give reason to re-think notions of fairness in terms of mid- and long-term outcomes and reconsider demographic parity as a helpful intervention whenever decision making systems are embedded in vicious cycles. Otherwise, even well-intended and well-constructed systems may stabilize and exacerbate inequality.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *Pro Publica*, 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. In *Tutorial at ICML 2018*, 2018.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, 2018. URL `https://uk.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secre`

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCS 2012*, pages 214–226, 2012. doi:10.1145/2090236.2090255.

European Union Agency for Fundamental Rights. *Article 12, Non-Discrimination*, volume C 326/391. 2012. URL `https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:12012P/TXT&from=DE`.

Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS 2016 Workship "ML and the Law"*, 2016. URL http://www.mlandthelaw.org/papers/grgic.pdf.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NIPS 2016*, pages 3315–3323, 2016. URL http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.

Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *WWW 2018*, pages 1389–1398, 2018. doi:10.1145/3178876.3186044.

Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *NIPS 2017*, pages 656–666. 2017. URL http://papers.nips.cc/paper/6668-avoiding-discrimination-through-causal-reasoning.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NIPS 2017*, pages 4066–4076, 2017. URL http://papers.nips.cc/paper/6995-counterfactual-fairness.

Lydia Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer Dy and Andreas Krause, editors, *ICML 2018*, pages 3156–3164, 2018. URL http://proceedings.mlr.press/v80/liu18c.html.

Cecilia Munoz, Megan Smith, and DJ Patil. Big data: A report on algorithmic systems, opportunity, and civil rights, 2016. URL https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_d

Rachel O'Dwyer. Algorithms are making the same mistakes assessing credit scores that humans did a century ago. *Quartz*, 2018. URL https://qz.com/1276781/algorithms-are-making-the-same-mistakes-assessing-credit-scores-t

Cathy O'Neil. *Weapons of Math Destruction: How big data increases inequality and threatens democracy*. Crown, Random House, New York City, NY, USA, 2016. ISBN 978-0553418811. URL https://weaponsofmathdestructionbook.com/.

Benjamin Paaßen, Astrid Bunge, Carolin Hainke, Leon Sindelar, and Matthias Vogelsang. Dynamic fairness – breaking vicious cycles in automatic decision making. In Michel Verleysen, editor, *Proceedings of the 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2019)*, 2019. accepted.

Indrė Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089, 2017. doi:10.1007/s10618-017-0506-1.

A Generalized Setup

In this appendix, we generalize our argument from the main paper to general probability densities. Further, we perform the stability analysis in more detail for three common probability distributions, namely the exponential distribution (Section A.4), the Pareto distribution (Section A.5), and the Gaussian distribution (Section A.6).

Our generalized setup is as follows. We assume a population with m individuals, a subset of which belong to a protected group $C \subseteq \{1, \ldots, m\}$. We denote the size of the protected group as $m_c = |C|$ and the size of the non-protected group as $m_{\neg c} = m - m_c$. We generally assume that $0 < m_c < m_{\neg c} < m$.

Further, we assume that every individual $i \in \{1, \ldots, m\}$ has an objective risk score q_i^t at time t, which is a real-valued random variable that is distributed according to some density p_c^t if $i \in C$ or according to some density $p_{\neg c}^t$ if $i \in \neg C$. Note that these densities may change over time and are thus indexed with the time step t. Further, we assume that an individual i is successful, i.e. $y_i = 1$, if and only if q_i^t is among the top n scores at time t. We also assume that $0 < n \ll m_c$, i.e. an acceptance is rare.

Under these assumptions, the following classifier $f^t : \{1, \ldots, m\} \rightarrow \{0, 1\}$ is per construction perfectly accurate.

$$
\begin{aligned}
f^t(i) &= \begin{cases} 1 & \text{if } g^t(i) \geq \theta^t \\ 0 & \text{otherwise} \end{cases} && \text{where} \\
g^t(i) &= q_i^t \\
\theta^t \text{ s.t.} &\quad |\{i | q_i^t \geq \theta^t\}| = n
\end{aligned}
$$

In other words, we predict success for individual i at time t by first assigning the risk score q_i^t and then applying a threshold that only leaves the top n people, i.e. exactly those who actually will be successful.

Note that this classifier is not only perfectly accurate but also conforms to equalized odds because we do not misclassify anyone, such that the rate of misclassifications is equal among all groups [Hardt et al., 2016]. Further, the scoring function is calibrated, the classifier is as accurate as possible, and it applies the same threshold for everyone, such that the fairness constraints of Corbett-Davies and Goel [2018] are fulfilled. Finally, the classifier fulfills notions of due process because it only accesses the objective risk score and makes no use of any other features of the individual [O'Neil, 2016, Dwork et al., 2012, Grgić-Hlača et al., 2016, Kilbertus et al., 2017, Kusner et al., 2017].

Next, we consider the probability of a person inside our outside the protected group to be classified as successful at time t, which is both equivalent to the probability of being successful at time t and to the probability of having a score above or equal to the threshold θ^t at time t. We denote this probability as P_c^t for the protected group and $P_{\neg c}^t$ otherwise. These probabilities are given as follows.

$$
\begin{aligned}
P_c^t &= \int_{\theta^t}^{\infty} p_c^t(q) dq \\
P_c^t &= \int_{\theta^t}^{\infty} p_c^t(q) dq
\end{aligned}
$$

Given these probabilities, we can estimate the number of people inside and outside the protected group who will be successful at time t. For each individual, success at time t is an independent Bernoulli trial with success probability P_c^t or $P_{\neg c}^t$ respectively. Accordingly, the sums $\sum_{i \in C} f(i)$ and $\sum_{i \in \neg C} f(i)$ are binomially distributed random variables with means

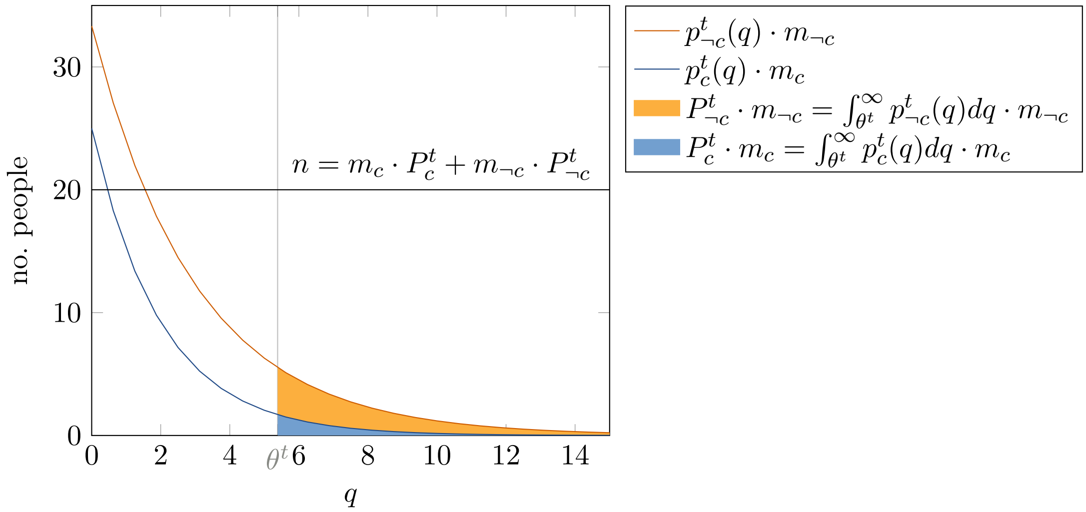

Figure 2: An illustration of Equation 2 for $n = 20$, $m_c = 50$, $m_{\neg c} = 100$, $\mu^t_c = 2$, and $\mu^t_{\neg c} = 3$. The threshold θ^t is selected on the x-axis such that exactly the amount of probability mass from both p^t_c and $p^t_{\neg c}$ lies on the right side of θ^t so that $m_c \cdot P^t_c$ and $m_{\neg c} \cdot P^t_{\neg c}$ add up to m.

$m_c \cdot P^t_c$ as well as $m_{\neg c} \cdot P^t_{\neg c}$ and variances $m_c \cdot P^t_c \cdot (1 - P^t_c)$ as well as $m_{\neg c} \cdot P^t_{\neg c} \cdot (1 - P^t_{\neg c})$. Note that the probabilities P^t_c with $P^t_{\neg c}$ decrease for lower n. Therefore, our assumption $n \ll m_c < m_{\neg c}$ implies that the variance is close to zero and thus the mean is a precise estimate of the actual number of successful people.

Using our mean estimate for the number of successful people in each group, the threshold θ^t can be estimated using the approximate equation

$$n = m_c \cdot P^t_c + m_{\neg c} \cdot P^t_{\neg c}. \tag{2}$$

Figuratively speaking, we slide θ^t from right to left along the score axis until we have collected enough probability mass from both p^t_c and $p^t_{\neg c}$ such that exactly n people are expected to have a score above the threshold (also refer to Figure 2).

A.1 Dynamical Model

Finally, we model the dynamics of our system. Our central modeling decision is that the means of the densities p^t_c and $p^t_{\neg c}$ shift over time. More precisely, we denote the means at time t as μ^t_c and $\mu^t_{\neg c}$, and assume the following dynamical system equation.

$$\begin{pmatrix} \mu^{t+1}_c \\ \mu^{t+1}_{\neg c} \end{pmatrix} = (1 - \alpha) \cdot \begin{pmatrix} \mu^t_c \\ \mu^t_{\neg c} \end{pmatrix} + \beta \cdot \begin{pmatrix} P^t_c \\ P^t_{\neg c} \end{pmatrix} \tag{3}$$

where $\alpha \in [0, 1]$ is a hyper-parameter describing how much score an individual loses in each time step ("leak reate"), and where $\beta \in \mathbb{R}^+$ is a hyper-parameter describing how much score an individual gains for for each positive classifier decision. By averaging these scores inside and outside the protected group, we obtain Equation 3. Note that the dynamics of μ^t_c and $\mu^t_{\neg c}$ are implicitly coupled due to the shared threshold θ^t.

Now, let us analyze the dynamic behavior of this system. In particular, we consider the score gap between the protected group and everyone else, which we denote as $\eta^t := \mu^t_{\neg c} - \mu^t_c$.

This score gap grows at time step t exactly if

$$\begin{aligned} & & \eta^{t+1} - \eta^t > 0 \\ \iff & & \mu_{\neg c}^{t+1} - \mu_c^{t+1} - \mu_{\neg c}^t + \mu_c^t > 0 \\ \iff & & (1-\alpha) \cdot \mu_{\neg c}^t + \beta \cdot P_{\neg c}^t - (1-\alpha) \cdot \mu_c^t - \beta \cdot P_c^t - \mu_{\neg c}^t + \mu_c^t > 0 \\ \iff & & \beta \cdot \left(P_{\neg c}^t - P_c^t\right) > \alpha \cdot \eta^t. \end{aligned} \tag{4}$$

Conversely, the score gap shrinks at time step at t exactly if

$$\beta \cdot \left(P_{\neg c}^t - P_c^t\right) < \alpha \cdot \eta^t. \tag{5}$$

From these equations, we can infer that the absolute value of the score gap will grow if α is sufficiently small, β is sufficiently large, and the probability distribution emphasizes gaps in the mean, i.e. small differences between means imply larger differences in probability mass on the margins. This covers a broad range of distributions where probability mass is concentrated around the mean.

A.2 Equilibria

Now, let us analyze the equilibria of our dynamical system. First, let us consider undesirable equilibria with a nonzero score gap. In particular, let us assume that the protected group is entirely unsuccessful, i.e. $P_c^* = 0$. Then, by virtue of Equation 2, we obtain $P_{\neg c}^* = \frac{n}{m_{\neg c}}$. The fix point equation yields:

$$\begin{pmatrix} \mu_c^* \\ \mu_{\neg c}^* \end{pmatrix} = (1-\alpha) \cdot \begin{pmatrix} \mu_c^* \\ \mu_{\neg c}^* \end{pmatrix} + \beta \cdot \begin{pmatrix} 0 \\ \frac{n}{m_{\neg c}} \end{pmatrix} \quad \iff \quad \begin{pmatrix} \mu_c^* \\ \mu_{\neg c}^* \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{\beta}{\alpha} \cdot \frac{n}{m_{\neg c}} \end{pmatrix}$$

In other words, we achieve an equilibrium if the protected group has a mean score of zero and no success whereas everybody else has a mean score of $\frac{\beta}{\alpha} \cdot \frac{n}{m_{\neg c}}$ and success probability $\frac{n}{m_{\neg c}}$. For many probability distributions, we can assume that probabilities remain unaffected by small deviations in the score gap. Therefore, we can demonstrate stability as follows.

First, assume that the score gap is slightly larger compared to the equilibrium. In this case, we obtain:

$$\alpha \cdot \eta^t > \alpha \cdot (\mu_{\neg c}^* - \mu_c^*) = \alpha \cdot \frac{\beta}{\alpha} \cdot \frac{n}{m_{\neg c}} = \beta \cdot \frac{n}{m_{\neg c}} = \beta \cdot (P_{\neg c}^* - P_c^*) \approx \beta \cdot (P_{\neg c}^t - P_c^t)$$

which in turn implies by virtue of Equation 5 that the score gap will shrink again. Second, assume that the score gap is slightly smaller compared to the equilibrium. In this case, we obtain:

$$\alpha \cdot \eta^t < \alpha \cdot (\mu_{\neg c}^* - \mu_c^*) = \alpha \cdot \frac{\beta}{\alpha} \cdot \frac{n}{m_{\neg c}} = \beta \cdot \frac{n}{m_{\neg c}} = \beta \cdot (P_{\neg c}^* - P_c^*) \approx \beta \cdot (P_{\neg c}^t - P_c^t)$$

which in turn implies by virtue of Equation 4 that the score gap grows again. Overall, the dynamical system counteracts small deviations, implying stability.

Next, we consider the desirable case, where success probabilities become equal across groups, i.e. $P_c^* = P_{\neg c}^* = P^*$. Then, Equation 2 yields $n = P^* \cdot m_c + P^* \cdot m_{\neg c}$, which implies $P^* = \frac{n}{m}$. Accordingly, the fix point equation yields:

$$\begin{pmatrix} \mu_c^* \\ \mu_{\neg c}^* \end{pmatrix} = (1-\alpha) \cdot \begin{pmatrix} \mu_c^* \\ \mu_{\neg c}^* \end{pmatrix} + \beta \cdot \begin{pmatrix} P^* \\ P^* \end{pmatrix} \quad \iff \quad \begin{pmatrix} \mu_c^* \\ \mu_{\neg c}^* \end{pmatrix} = \begin{pmatrix} \frac{\beta}{\alpha} \cdot \frac{n}{m} \\ \frac{\beta}{\alpha} \cdot \frac{n}{m} \end{pmatrix}$$

In other words, the mean for both groups has the same value $\mu^*_c = \mu^*_{\neg c} = \frac{\beta}{\alpha} \cdot \frac{n}{m}$. Unfortunately, this point is not stable in general, because a slight nonzero score gap will, for many probability distributions, result in emphasized gaps in success probability, which in turn may fulfill Equation 4 or 5, leading to even more pronounced score gaps (as we will show in Sections A.4, A.5, and A.6). To counteract this potential instability, we can employ demographic parity.

A.3 Demographic Parity

Demographic parity requires that, at any time step t, $P^t_c = P^t_{\neg c}$, which we can re-write to $P^t_{\neg c} - P^t_c = 0$. This, in turn, implies that Equation 4 is always fulfilled if the score gap is negative and Equation 5 is always fulfilled if the score gap is positive. In other words, demographic parity ensures that only states with zero score gaps can be equilibria and that these equilibria are always stable. In case there is a one-to-one map from means to success probabilities (as is the case in most common probability distributions), this in turn implies that there is a unique, stable equilibrium where the score gap and the probability gap are both zero, i.e. $P^*_c = P^*_{\neg c} = \frac{n}{m}$ and $\mu^*_c = \mu^*_{\neg c} = \frac{\beta}{\alpha} \cdot \frac{n}{m}$.

We can also confirm the stability finding using classic stability theory. If $P^t_c = P^t_{\neg c}$, then Equation 2 implies that $P^t_c = P^t_{\neg c} = \frac{n}{m}$ for all time steps. Therefore, we obtain the following Jacobian for Equation 3

$$J(\mu^t_c, \mu^t_{\neg c}) = \begin{pmatrix} \frac{\partial}{\partial \mu^t_c}\big[(1-\alpha)\cdot\mu^t_c + \beta\cdot\frac{n}{m}\big] & \frac{\partial}{\partial \mu^t_{\neg c}}\big[(1-\alpha)\cdot\mu^t_c + \beta\cdot\frac{n}{m}\big] \\ \frac{\partial}{\partial \mu^t_c}\big[(1-\alpha)\cdot\mu^t_{\neg c} + \beta\cdot\frac{n}{m}\big] & \frac{\partial}{\partial \mu^t_{\neg c}}\big[(1-\alpha)\cdot\mu^t_{\neg c} + \beta\cdot\frac{n}{m}\big] \end{pmatrix} = \begin{pmatrix} 1-\alpha & 0 \\ 0 & 1-\alpha \end{pmatrix}$$

The two eigenvalues of this matrix are both $1 - \alpha$. Therefore, for a nonzero leak rate, the absolute value of these eigenvalues is smaller than one and therefore stability theory implies that the equilibrium is stable. Note that this result applies independent of the probability distribution in question and relies only on a one-to-one mapping between means and success probabilities.

In the following sections, we consider the stability findings from the previous section in more detail for three specific distributions, namely the exponential, the Pareto, and the Gaussian distribution.

A.4 Stability for the Exponential Distribution

In the exponential distribution model, we assume that the densities p^t_c and $p^t_{\neg c}$ have the following form (also refer to Figure 2).

$$p^t_c(q) = \frac{1}{\mu^t_c} \cdot \exp\big(-\frac{q}{\mu^t_c}\big)$$
$$p^t_{\neg c}(q) = \frac{1}{\mu^t_{\neg c}} \cdot \exp\big(-\frac{q}{\mu^t_{\neg c}}\big)$$

Note that the densities are fully parameterized by the respective mean, which makes the exponential distribution a straightforward object of study. For these densities, the probabilities P^t_c and $P^t_{\neg c}$ take the following form.

$$P^t_c = \int_{\theta^t}^{\infty} p^t_c(q)dq = \exp\big(-\frac{\theta^t}{\mu^t_c}\big)$$
$$P^t_{\neg c} = \int_{\theta^t}^{\infty} p^t_{\neg c}(q)dq = \exp\big(-\frac{\theta^t}{\mu^t_{\neg c}}\big)$$

Now, let us consider the undesirable equilibrium case, where $\mu^*_c = 0$ and $\mu^*_{\neg c} = \frac{\beta}{\alpha} \cdot \frac{n}{m_{\neg c}}$. First, note that the threshold θ^t in this condition is lower-bounded due to Equation 2. In particular, we obtain the following lower bound.

$$\begin{aligned} && \exp\big(-\frac{\theta^t}{\mu^*_{\neg c}}\big) = \exp\big(-\theta^t \cdot \frac{\alpha}{\beta} \cdot \frac{m_{\neg c}}{n}\big) &\le \frac{n}{m_{\neg c}} \\ \iff && -\theta^t \cdot \frac{\alpha}{\beta} \cdot \frac{m_{\neg c}}{n} &\le \log\big(\frac{n}{m_{\neg c}}\big) \\ \iff && \theta^t &\ge -\frac{\beta}{\alpha} \cdot \frac{n}{m_{\neg c}} \log\big(\frac{n}{m_{\neg c}}\big) \end{aligned}$$

Note that this term is strictly larger than zero as $n \ll m_{\neg c}$, which implies that $\log\big(\frac{n}{m_{\neg c}}\big) < 0$. Accordingly, if μ^t_c is sufficiently small, i.e. $\mu^t_c \ll -\frac{\beta}{\alpha} \cdot \frac{n}{m_{\neg c}} \log\big(\frac{n}{m_{\neg c}}\big)$, then $P^t_c = \exp\big(-\frac{\theta^t}{\mu^t_c}\big) \approx 0$. Also note that small changes in μ^t_c will not change the probability P^t_c in this case. Therefore, for sufficiently small μ^t_c, we obtain approximatively constant probabilities $P^t_c \approx 0$ and $P^t_{\neg c} \approx \frac{n}{m_{\neg c}}$. Accordingly, the Jacobian matrix of Equation 3 for sufficiently small μ^t_c is given as follows.

$$J(\mu^t_c, \mu^t_{\neg c}) = \begin{pmatrix} \frac{\partial}{\partial \mu^t_c}\big[(1-\alpha)\cdot\mu^t_c + \beta \cdot 0\big] & \frac{\partial}{\partial \mu^t_{\neg c}}\big[(1-\alpha)\cdot\mu^t_c + \beta \cdot 0\big] \\ \frac{\partial}{\partial \mu^t_c}\big[(1-\alpha)\cdot\mu^t_{\neg c} + \beta \cdot \frac{n}{m_{\neg c}}\big] & \frac{\partial}{\partial \mu^t_{\neg c}}\big[(1-\alpha)\cdot\mu^t_{\neg c} + \beta \cdot \frac{n}{m_{\neg c}}\big] \end{pmatrix} = \begin{pmatrix} 1-\alpha & 0 \\ 0 & 1-\alpha \end{pmatrix}$$

The two eigenvalues of this matrix are both $1 - \alpha$. Therefore, for a nonzero leak rate, the absolute value of these eigenvalues is smaller than one and therefore stability theory implies that the equilibrium is stable.

Next, let us consider the desirable equilibrium case, where $\mu^*_c = \mu^*_{\neg c} = \frac{\beta}{\alpha} \cdot \frac{n}{m}$ and $P^*_c = P^*_{\neg c} = \frac{n}{m}$. In this equilibrium condition, we can obtain the threshold θ^* as follows.

$$\begin{aligned} && \frac{n}{m} = P^*_c &= \exp\big(-\frac{\theta^*}{\mu^*_c}\big) \\ \iff && \log\big(\frac{n}{m}\big) &= -\frac{\theta^*}{\mu^*_c} \\ \iff && \theta^* &= -\mu^*_c \cdot \log\big(\frac{n}{m}\big) \end{aligned}$$

Now, let us consider small deviations of μ^t_c and $\mu^t_{\neg c}$ such that θ^t stays equal to θ^*. Such deviations are possible since we can let both means deviate in opposite directions such that the condition holds. In that case, we obtain the following Jacobian of our model in Equation 3.

$$\begin{aligned} J(\mu^t_c, \mu^t_{\neg c}) &= \begin{pmatrix} \frac{\partial}{\partial \mu^t_c}\big[(1-\alpha)\cdot\mu^t_c + \beta \cdot \exp\big(-\frac{\theta^*}{\mu^t_c}\big)\big] & \frac{\partial}{\partial \mu^t_{\neg c}}\big[(1-\alpha)\cdot\mu^t_c + \beta \cdot \exp\big(-\frac{\theta^*}{\mu^t_c}\big)\big] \\ \frac{\partial}{\partial \mu^t_c}\big[(1-\alpha)\cdot\mu^t_{\neg c} + \beta \cdot \exp\big(-\frac{\theta^*}{\mu^t_{\neg c}}\big)\big] & \frac{\partial}{\partial \mu^t_{\neg c}}\big[(1-\alpha)\cdot\mu^t_{\neg c} + \beta \cdot \exp\big(-\frac{\theta^*}{\mu^t_{\neg c}}\big)\big] \end{pmatrix} \\ &= \begin{pmatrix} 1-\alpha+\beta\cdot\exp\big(-\frac{\theta^*}{\mu^t_c}\big)\cdot\frac{\theta^*}{(\mu^t_c)^2} & 0 \\ 0 & 1-\alpha+\beta\cdot\exp\big(-\frac{\theta^*}{\mu^t_{\neg c}}\big)\cdot\frac{\theta^*}{(\mu^t_{\neg c})^2} \end{pmatrix} \end{aligned}$$

Accordingly, the Jacobian at our equilibrium is given as follows.

$$\begin{aligned} J(\mu^*_c, \mu^*_{\neg c}) &= \begin{pmatrix} 1-\alpha+\beta\cdot\exp\big(-\frac{\theta^*}{\mu^*_c}\big)\cdot\frac{\theta^*}{(\mu^*_c)^2} & 0 \\ 0 & 1-\alpha+\beta\cdot\exp\big(-\frac{\theta^*}{\mu^*_{\neg c}}\big)\cdot\frac{\theta^*}{(\mu^*_{\neg c})^2} \end{pmatrix} \\ &= \begin{pmatrix} 1-\alpha-\beta\cdot\exp\big(\log\big(\frac{n}{m}\big)\big)\cdot\frac{\alpha}{\beta}\cdot\frac{m}{n}\cdot\log\big(\frac{n}{m}\big) & 0 \\ 0 & 1-\alpha-\beta\cdot\exp\big(\log\big(\frac{n}{m}\big)\big)\cdot\frac{\alpha}{\beta}\cdot\frac{m}{n}\cdot\log\big(\frac{n}{m}\big) \end{pmatrix} \\ &= \begin{pmatrix} 1-\alpha-\alpha\cdot\log\big(\frac{n}{m}\big) & 0 \\ 0 & 1-\alpha-\alpha\cdot\log\big(\frac{n}{m}\big) \end{pmatrix} \end{aligned}$$

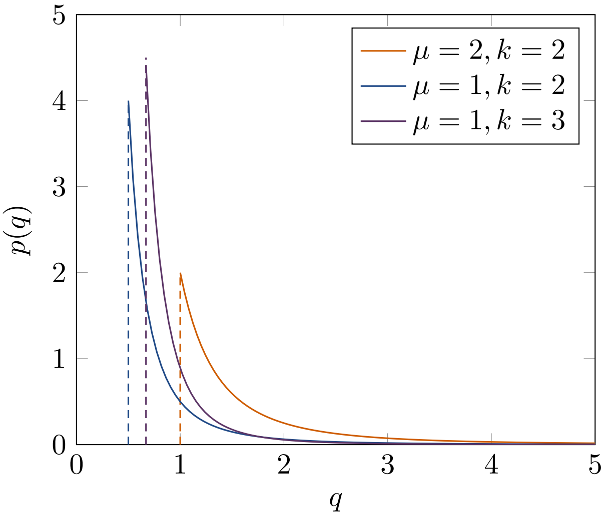

Figure 3: Three different Pareto densities with means μ and rate parameters k as specified in the legend.

The two eigenvalues of this Jacobian are both $1-\alpha-\alpha\cdot\log\left(\frac{n}{m}\right)$. Accordingly, our equilibrium is unstable if $|1-\alpha-\alpha\cdot\log\left(\frac{n}{m}\right)| > 1$. Given that $n \ll m$ and $\alpha \in [0,1]$, $1-\alpha-\alpha\cdot\log\left(\frac{n}{m}\right)$ is larger than 0. Therefore, we can re-write the instability condition as follows.

$$
\begin{aligned}
& 1-\alpha-\alpha\cdot\log\left(\frac{n}{m}\right) > 1 \\
\iff\quad & 1+\log\left(\frac{n}{m}\right) < 0 \\
\iff\quad & \frac{n}{m} < \frac{1}{e}
\end{aligned}
$$

In other words, if m is at least e times larger than n, the equilibrium is unstable. From $n \ll m_c < \frac{1}{2}m$, we can conclude that this is the case.

In summary, we have demonstrated that the exponential distribution yields attractive equilibria in undesirable positions, whereas the desirable equilibria are unstable for a wide range of conditions.

A.5 Stability for the Pareto Distribution

In the Pareto distribution model, we assume that the densities p^t_c and $p^t_{\neg c}$ have the following form (also refer to Figure 3).

$$
\begin{aligned}
p^t_c(q) &= \frac{(k-1)^k}{k^{k-1}}\cdot\frac{(\mu^t_c)^k}{q^{k+1}} \\
p^t_{\neg c}(q) &= \frac{(k-1)^k}{k^{k-1}}\cdot\frac{(\mu^t_{\neg c})^k}{q^{k+1}}
\end{aligned}
$$

where $k > 1$ controls the slope of the distribution and how much probability mass is concentrated around the mean. Note that the Pareto distribution is only defined on the interval $[\frac{k-1}{k}\cdot\mu, \infty)$ and that the variance is infinite for $k \le 2$.

From these densities we obtain the following success probabilities P_c^t and $P_{\neg c}^t$.

$$P_c^t = \int_{\theta^t}^{\infty} p_c^t(q) dq = \Big(\frac{k-1}{k}\Big)^k \cdot \Big(\frac{\mu_c^t}{\theta^t}\Big)^k$$
$$P_{\neg c}^t = \int_{\theta^t}^{\infty} p_{\neg c}^t(q) dq = \Big(\frac{k-1}{k}\Big)^k \cdot \Big(\frac{\mu_{\neg c}^t}{\theta^t}\Big)^k$$

Plugging these results into Equation 2, we obtain a closed-form expression for the threshold as follows.

$$\begin{aligned} & n = m_c \cdot \Big(\frac{k-1}{k}\Big)^k \cdot \Big(\frac{\mu_c^t}{\theta^t}\Big)^k + m_{\neg c} \cdot \Big(\frac{k-1}{k}\Big)^k \cdot \Big(\frac{\mu_{\neg c}^t}{\theta^t}\Big)^k \\ \iff \quad & n \cdot (\theta^t)^k = \Big(\frac{k-1}{k}\Big)^k \cdot \big[m_c \cdot (\mu_c^t)^k + m_{\neg c} \cdot (\mu_{\neg c}^t)^k\big] \\ \iff \quad & \theta^t = \frac{k-1}{k} \cdot \sqrt[k]{\frac{1}{n}\big[m_c \cdot (\mu_c^t)^k + m_{\neg c} \cdot (\mu_{\neg c}^t)^k\big]} \end{aligned}$$

Next, we plug this solution back into our probability expressions and obtain:

$$P_c^t = \frac{n \cdot (\mu_c^t)^k}{m_c \cdot (\mu_c^t)^k + m_{\neg c} \cdot (\mu_{\neg c}^t)^k}$$
$$P_{\neg c}^t = \frac{n \cdot (\mu_{\neg c}^t)^k}{m_c \cdot (\mu_c^t)^k + m_{\neg c} \cdot (\mu_{\neg c}^t)^k}$$

Accordingly, we obtain the following Jacobian.

$$\begin{aligned} & J(\mu_c^t, \mu_{\neg c}^t) \\ = & \begin{pmatrix} \frac{\partial}{\partial \mu_c^t}\Big[(1-\alpha)\cdot\mu_c^t + \beta \cdot \frac{n\cdot(\mu_c^t)^k}{m_c\cdot(\mu_c^t)^k + m_{\neg c}\cdot(\mu_{\neg c}^t)^k}\Big] & \frac{\partial}{\partial \mu_{\neg c}^t}\Big[(1-\alpha)\cdot\mu_c^t + \beta \cdot \frac{n\cdot(\mu_c^t)^k}{m_c\cdot(\mu_c^t)^k + m_{\neg c}\cdot(\mu_{\neg c}^t)^k}\Big] \\ \frac{\partial}{\partial \mu_c^t}\Big[(1-\alpha)\cdot\mu_{\neg c}^t + \beta \frac{n\cdot(\mu_{\neg c}^t)^k}{m_c\cdot(\mu_c^t)^k + m_{\neg c}\cdot(\mu_{\neg c}^t)^k}\Big] & \frac{\partial}{\partial \mu_{\neg c}^t}\Big[(1-\alpha)\cdot\mu_{\neg c}^t + \beta \cdot \frac{n\cdot(\mu_{\neg c}^t)^k}{m_c\cdot(\mu_c^t)^k + m_{\neg c}\cdot(\mu_{\neg c}^t)^k}\Big] \end{pmatrix} \\ = & \begin{pmatrix} 1-\alpha+\beta\cdot k\cdot n\cdot(\mu_c^t)^{k-1}\cdot\frac{m_{\neg c}\cdot(\mu_{\neg c}^t)^k}{(m_c\cdot(\mu_c^t)^k + m_{\neg c}\cdot(\mu_{\neg c}^t)^k)^2} & -\beta\cdot\frac{n\cdot(\mu_c^t)^k}{(m_c\cdot(\mu_c^t)^k + m_{\neg c}\cdot(\mu_{\neg c}^t)^k)^2}\cdot m_{\neg c}\cdot k\cdot(\mu_{\neg c}^t)^{k-1} \\ -\beta\cdot\frac{n\cdot(\mu_{\neg c}^t)^k}{(m_c\cdot(\mu_c^t)^k + m_{\neg c}\cdot(\mu_{\neg c}^t)^k)^2}\cdot m_c\cdot k\cdot(\mu_c^t)^{k-1} & 1-\alpha+\beta\cdot k\cdot n\cdot(\mu_{\neg c}^t)^{k-1}\cdot\frac{m_c\cdot(\mu_c^t)^k}{(m_c\cdot(\mu_c^t)^k + m_{\neg c}\cdot(\mu_{\neg c}^t)^k)^2} \end{pmatrix} \\ = & \begin{pmatrix} 1-\alpha+\frac{\beta\cdot k\cdot(\mu_c^t)^{k-1}}{m_c\cdot(\mu_c^t)^k + m_{\neg c}\cdot(\mu_{\neg c}^t)^k}\cdot P_{\neg c}^t & -\frac{\beta\cdot m_{\neg c}\cdot k\cdot(\mu_{\neg c}^t)^{k-1}}{m_c\cdot(\mu_c^t)^k + m_{\neg c}\cdot(\mu_{\neg c}^t)^k}\cdot P_c^t \\ -\frac{\beta\cdot m_c\cdot k\cdot(\mu_c^t)^{k-1}}{m_c\cdot(\mu_c^t)^k + m_{\neg c}\cdot(\mu_{\neg c}^t)^k}\cdot P_{\neg c}^t & 1-\alpha+\frac{\beta\cdot k\cdot(\mu_{\neg c}^t)^{k-1}}{m_c\cdot(\mu_c^t)^k + m_{\neg c}\cdot(\mu_{\neg c}^t)^k}\cdot P_c^t \end{pmatrix} \end{aligned}$$

Now, let us consider the undesirable equilibrium case, where $\mu_c^* = \epsilon$ with $\epsilon \approx 0$[3], $\mu_{\neg c}^* = \frac{\beta}{\alpha} \cdot \frac{n}{m_{\neg c}}$, $P_c^* = 0$, and $P_{\neg c}^* = \frac{n}{m_{\neg c}}$. For this equilibrium, we obtain the following Jacobian.

$$J(\epsilon, \frac{\beta}{\alpha} \cdot \frac{n}{m_{\neg c}}) \approx \begin{pmatrix} 1-\alpha & 0 \\ 0 & 1-\alpha \end{pmatrix}$$

The two eigenvalues of this matrix are both $1-\alpha$. Therefore, for a nonzero leak rate, the absolute value of these eigenvalues is smaller than one and therefore stability theory implies that the equilibrium is stable.

[3]A mean of zero would imply an ill-defined Pareto density. Therefore, we consider here a mean close to zero.

Next, let us consider the desirable equilibrium case, where $\mu^* = \mu^*_c = \mu^*_{\neg c} = \frac{\beta}{\alpha} \cdot \frac{n}{m}$ and $P^* = P^*_c = P^*_{\neg c} = \frac{n}{m}$. For this equilibrium, we obtain the following Jacobian.

$$J(\frac{\beta}{\alpha} \cdot \frac{n}{m}, \frac{\beta}{\alpha} \cdot \frac{n}{m}) = \begin{pmatrix} 1 - \alpha + \frac{\beta \cdot k \cdot (\mu^*)^{k-1}}{m \cdot (\mu^*)^k} \cdot P^* & -\frac{\beta \cdot m_{\neg c} \cdot k \cdot (\mu^*)^{k-1}}{m \cdot (\mu^*)^k} \cdot P^* \\ -\frac{\beta \cdot m_c \cdot k \cdot (\mu^*)^{k-1}}{m \cdot (\mu^*)^k} \cdot P^* & 1 - \alpha + \frac{\beta \cdot k \cdot (\mu^*)^{k-1}}{m \cdot (\mu^*)^k} \cdot P^* \end{pmatrix}$$
$$= \begin{pmatrix} 1 - \alpha + \frac{\beta \cdot k}{m} \cdot \frac{P^*}{\mu^*} & -\beta \cdot k \cdot \frac{m_{\neg c}}{m} \cdot \frac{P^*}{\mu^*} \\ -\beta \cdot k \cdot \frac{m_c}{m} \cdot \frac{P^*}{\mu^*} & 1 - \alpha + \frac{\beta \cdot k}{m} \cdot \frac{P^*}{\mu^*} \end{pmatrix}$$
$$= \begin{pmatrix} 1 - \alpha \cdot \left[1 + \frac{k}{m}\right] & -\alpha \cdot k \cdot \frac{m_{\neg c}}{m} \\ -\alpha \cdot k \cdot \frac{m_c}{m} & 1 - \alpha \cdot \left[1 + \frac{k}{m}\right] \end{pmatrix}$$

The eigenvalues of this Jacobian are $\lambda_1 = 1 - \alpha \cdot \left[1 + \frac{k}{m} \cdot (1 + \sqrt{m_c \cdot m_{\neg c}})\right]$ and $\lambda_1 = 1 - \alpha \cdot \left[1 + \frac{k}{m} \cdot (1 + \sqrt{m_c \cdot m_{\neg c}})\right]$. The absolute values of these eigenvalues exceed 1 if the following respective conditions hold.

$$\lambda_1 > 1 \qquad \iff \qquad k > \frac{m}{\sqrt{m_c \cdot m_{\neg c}} - 1}$$
$$\lambda_2 < -1 \qquad \iff \qquad k > \left(\frac{2}{\alpha} - 1\right) \cdot \frac{m}{\sqrt{m_c \cdot m_{\neg c}} + 1}$$

The former condition can be fulfilled easily if $k > 2$ and $m_c \approx m_{\neg c} \approx \frac{m}{2}$. As such, this equilibrium is unstable for a wide range of possible settings.

In summary, we have demonstrated that the Pareto distribution yields attractive equilibria in undesirable positions, whereas the desirable equilibria are unstable for a wide range of conditions.

A.6 Stability for the Gaussian Distribution

In the Gaussian distribution model, we assume that the densities p^t_c and $p^t_{\neg c}$ have the following form:

$$p^t_c(q) = \mathcal{N}(q|\mu^t_c, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2} \cdot \frac{(q - \mu^t_c)^2}{\sigma^2}\right)$$
$$p^t_{\neg c}(q) = \mathcal{N}(q|\mu^t_{\neg c}, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2} \cdot \frac{(q - \mu^t_{\neg c})^2}{\sigma^2}\right)$$

where $\mathcal{N}$ denotes the Gaussian density function and where σ is the standard deviation of the Gaussian, which we assume to be fixed and equal across groups.

From these densities we obtain the following success probabilities P^t_c and $P^t_{\neg c}$.

$$P^t_c = \int_{\theta^t}^{\infty} p^t_c(q)dq = 1 - \Phi\left(\frac{\theta^t - \mu^t_c}{\sigma}\right)$$
$$P^t_{\neg c} = \int_{\theta^t}^{\infty} p^t_{\neg c}(q)dq = 1 - \Phi\left(\frac{\theta^t - \mu^t_{\neg c}}{\sigma}\right)$$

where Φ is the cumulative density function of the standard Gaussian distribution.

Now, let us consider the undesirable equilibrium case, where $\mu^*_c = 0$ and $\mu^*_{\neg c} = \frac{\beta}{\alpha} \cdot \frac{n}{m_{\neg c}}$. First, note that the threshold θ^t is lower-bounded in this condition due to Equation 2. In particular, we obtain the following lower bound.

$$
\begin{aligned}
& 1 - \Phi\big(\frac{\theta^t - \mu^*_{\neg c}}{\sigma}\big) \leq \frac{n}{m_{\neg c}} \\
\iff \quad & \frac{\theta^t - \mu^*_{\neg c}}{\sigma} \geq \Phi^{-1}\big(1 - \frac{n}{m_{\neg c}}\big) \\
\iff \quad & \theta^t \geq \sigma \cdot \Phi^{-1}\big(1 - \frac{n}{m_{\neg c}}\big) + \mu^*_{\neg c}
\end{aligned}
$$

Note that this term is strictly larger than zero. Accordingly, for sufficiently small σ and small $\mu^t_c < \mu^*_{\neg c}$ we obtain:

$$
P^t_c = 1 - \Phi\Big(\frac{\theta^t - \mu^t_c}{\sigma}\Big) \leq 1 - \Phi\Big(\Phi^{-1}\big(1 - \frac{n}{m_{\neg c}}\big) + \frac{\mu^*_{\neg c} - \mu^t_c}{\sigma}\Big) \leq 1 - \Phi\Big(\frac{\mu^*_{\neg c} - \mu^t_c}{\sigma}\Big) \approx 1 - 1 = 0
$$

In other words, for sufficiently small σ we obtain $P^t_c \approx P^*_c = 0$, even if we vary μ^t_c slightly. Accordingly, $P^t_{\neg c} = \frac{n - m_c \cdot P^*_c}{m_{\neg c}} \approx \frac{n}{m_{\neg c}} = P^*_{\neg c}$. This results in the following Jacobian matrix of Equation 3 for small μ^t_c, $\mu^t_{\neg c} \approx \mu^*_{\neg c}$ and small σ.

$$
J(\mu^t_c, \mu^t_{\neg c}) = \begin{pmatrix} \frac{\partial}{\partial \mu^t_c}\big[(1-\alpha)\cdot \mu^t_c + \beta \cdot 0\big] & \frac{\partial}{\partial \mu^t_{\neg c}}\big[(1-\alpha)\cdot \mu^t_c + \beta \cdot 0\big] \\ \frac{\partial}{\partial \mu^t_c}\big[(1-\alpha)\cdot \mu^t_{\neg c} + \beta \cdot \frac{n}{m_{\neg c}}\big] & \frac{\partial}{\partial \mu^t_{\neg c}}\big[(1-\alpha)\cdot \mu^t_{\neg c} + \beta \cdot \frac{n}{m_{\neg c}}\big] \end{pmatrix} = \begin{pmatrix} 1-\alpha & 0 \\ 0 & 1-\alpha \end{pmatrix}
$$

The two eigenvalues of this matrix are both $1 - \alpha$. Therefore, for a nonzero leak rate, the absolute value of these eigenvalues is smaller than one and therefore stability theory implies that the equilibrium is stable.

Next, let us consider the desirable equilibrium case, where $\mu^*_c = \mu^*_{\neg c} = \frac{\beta}{\alpha} \cdot \frac{n}{m}$ and $P^*_c = P^*_{\neg c} = \frac{n}{m}$. In this equilibrium condition, we can obtain the threshold θ^* as follows.

$$
\begin{aligned}
& \frac{n}{m} = P^*_c = 1 - \Phi\big(\frac{\theta^* - \mu^*_c}{\sigma}\big) \\
\iff \quad & \frac{\theta^* - \mu^*_c}{\sigma} = \Phi^{-1}\big(1 - \frac{n}{m}\big) \\
\iff \quad & \theta^* = \Phi^{-1}\big(1 - \frac{n}{m}\big) \cdot \sigma + \mu^*_c
\end{aligned}
$$

Now, let us consider small deviations of μ^t_c and $\mu^t_{\neg c}$ which are such that the threshold θ^t stays equal to θ^*. In that case, we obtain the following Jacobian of our model in Equation 3.

$$
\begin{aligned}
J(\mu^t_c, \mu^t_{\neg c}) &= \begin{pmatrix} \frac{\partial}{\partial \mu^t_c}\big[(1-\alpha)\cdot \mu^t_c + \beta \cdot \big(1 - \Phi(\frac{\theta^* - \mu^t_c}{\sigma})\big)\big] & \frac{\partial}{\partial \mu^t_{\neg c}}\big[(1-\alpha)\cdot \mu^t_c + \beta \cdot \big(1 - \Phi(\frac{\theta^* - \mu^t_c}{\sigma})\big)\big] \\ \frac{\partial}{\partial \mu^t_c}\big[(1-\alpha)\cdot \mu^t_{\neg c} + \beta \cdot \big(1 - \Phi(\frac{\theta^* - \mu^t_{\neg c}}{\sigma})\big)\big] & \frac{\partial}{\partial \mu^t_{\neg c}}\big[(1-\alpha)\cdot \mu^t_{\neg c} + \beta \cdot \big(1 - \Phi(\frac{\theta^* - \mu^t_{\neg c}}{\sigma})\big)\big] \end{pmatrix} \\
&= \begin{pmatrix} 1 - \alpha - \beta \cdot \mathcal{N}(\theta^* | \mu^t_c, \sigma) & 0 \\ 0 & 1 - \alpha - \beta \cdot \mathcal{N}(\theta^* | \mu^t_{\neg c}, \sigma) \end{pmatrix}
\end{aligned}
$$

Accordingly, the Jacobian at our equilibrium is given as follows.

$$
\begin{aligned}
J(\mu^*_c, \mu^*_{\neg c}) &= \begin{pmatrix} 1 - \alpha - \beta \cdot \mathcal{N}(\theta^* | \mu^*_c, \sigma) & 0 \\ 0 & 1 - \alpha - \beta \cdot \mathcal{N}(\theta^* | \mu^*_{\neg c}, \sigma) \end{pmatrix} \\
&= \begin{pmatrix} 1 - \alpha - \beta \cdot \mathcal{N}\Big(\Phi^{-1}\big(1 - \frac{n}{m}\big) \cdot \sigma \Big| 0, \sigma\Big) & 0 \\ 0 & 1 - \alpha - \beta \cdot \mathcal{N}\Big(\Phi^{-1}\big(1 - \frac{n}{m}\big) \cdot \sigma \Big| 0, \sigma\Big) \end{pmatrix} \\
&= \begin{pmatrix} 1 - \alpha - \frac{\beta}{\sigma} \cdot \mathcal{N}\Big(\Phi^{-1}\big(1 - \frac{n}{m}\big) \Big| 0, 1\Big) & 0 \\ 0 & 1 - \alpha - \frac{\beta}{\sigma} \cdot \mathcal{N}\Big(\Phi^{-1}\big(1 - \frac{n}{m}\big) \Big| 0, 1\Big) \end{pmatrix}
\end{aligned}
$$

The two eigenvalues of this Jacobian are both $1 - \alpha - \frac{\beta}{\sigma} \cdot \mathcal{N}\Big(\Phi^{-1}(1 - \frac{n}{m})\Big|0, 1\Big)$. Accordingly, our equilibrium is unstable if $|1 - \alpha - \frac{\beta}{\sigma} \cdot \mathcal{N}\Big(\Phi^{-1}(1 - \frac{n}{m})\Big|0, 1\Big)| > 1$. Given that α, β, and the Gaussian density function are all non-negative, we can re-write the instability condition as follows.

$$
\begin{aligned}
& & -1 > 1 - \alpha - \frac{\beta}{\sigma} \cdot \mathcal{N}\Big(\Phi^{-1}(1 - \frac{n}{m})\Big|0, 1\Big) \\
\iff & & \alpha - 2 > -\frac{\beta}{\sigma} \cdot \mathcal{N}\Big(\Phi^{-1}(1 - \frac{n}{m})\Big|0, 1\Big) \\
\iff & & \sigma < \frac{\beta}{2 - \alpha} \cdot \mathcal{N}\Big(\Phi^{-1}(1 - \frac{n}{m})\Big|0, 1\Big)
\end{aligned}
$$

In other words, for sufficiently small σ, the equilibrium is unstable.

In summary, we have demonstrated that the Gaussian distribution yields attractive equilibria in undesirable positions, whereas the desirable equilibria are unstable for a wide range of conditions.