Journal of Mechanism and Institution Design

CONTENTS

Volume 10, Issue 1, December 2025
ISSN: 2399-844X (Print), 2399-8458 (Online)

Editorial board

Published by
The Society for the Promotion of Mechanism and Institution Design
Editorial office, Centre for Mechanism and Institution Design
University of York, Heslington, York YO10 5DD
United Kingdom
https://www.mechanism-design.org

ISSN: 2399-844X (Print), 2399-8458 (Online), DOI: 10.22574/jmid

The founding institutional members are
University of York, UK
Universität Klagenfurt, Austria
Southwestern University of Economics and Finance, China.

Cover & Logo Artwork @ Jasmine Yang
LaTeX Editor & Journal Design @ Paul Schweinzer (based on ‘confproc’)

Printed in York and Klagenfurt by the Centre for Mechanism and Institution Design

A Letter from the Editors

THE world has experienced yet another year of turmoil in (inter)national politics, its institutions, trade, education, as well as individual study, visa, and work circumstances. The relative peace and stability of the previous half-century have given way to increased overall uncertainty as well as legal indeterminacy, resulting in a whole new set of critical, pressing questions for all of us, and, in particular, *The Anxious Generation*. Do we want to live in a world that is governed by the rule of law or by the law of the jungle?

Our already exciting environment has been further shocked by the arrival of a variety of increasingly capable large language models. Although cautioned against such practice, we use them to write our problem sets, our students use them to answer their assignments, and we have them graded using the same. The research evaluation and publication process will soon face similar practices. How we shall deal with these productively is a question this Journal will ask and, together with everyone else, will have to answer.

The *Society for the Promotion of Mechanism and Institution Design* will organize the 5$^{\text{th}}$ *Conference on Mechanism and Institution Design*, this time in York, UK, from June 22$^{\text{nd}}$ to 26$^{\text{th}}$, 2026. The conference will be organized by Yuan Ju, Paul Schweinzer, and Zaifu Yang and will be hosted by the University of York, Department of Economics and Related Studies. The CMID2026 keynote speakers will be Atila Abdulkadiroglu, Duke University, Yan Chen, University of Michigan, Matthew Jackson, Stanford University, and Olivier Tercieux, Paris School of Economics. Submissions are open now. We are looking forward to meeting participants in the Yorkshire summer sun.

In 2024, YingHua He, an associate professor of economics at Rice University, an outstanding scholar and a supporter of this Journal, died at the age of 44 after fighting a quickly spreading cancer. In order to commemorate Professor He's impactful career, this Journal will publish a special issue on market design in the coming year, inviting contributions that align with Professor He's diverse research interests. The guest editors are Yan Chen, University of Michigan, and Jeremy Fox, Rice University.

We are very pleased to report that Chia-Hui Chen, Maciej Kotowski, Jiangtao Li, Thanh Nguyen, and Alex Westkamp have recently joined our editorial board. We thank them and, as with each new issue, all editorial members, referees, and supporters for their contributions to the Journal. Everyone involved in the scientific selection and technical production of the Journal is volunteering their capabilities, time, and effort. As a result, our *Journal of Mechanism and Institution Design* can publish high-quality research at no cost to both authors and readers, providing free and open access to the public. We hope the community sharing this vision will grow, become stronger, and more successful in tackling the most pressing design issues of our time.

Paul Schweinzer & Zaifu Yang, Klagenfurt, Macau, & York, Dec 20th, 2025.

Journal of Mechanism and Institution Design
ISSN: 2399-844X (Print), 2399-8458 (Online)
DOI:10.22574/jmid.2025.12.001

MEDIATED (ANTI)PERSUASIVE COMMUNICATION

Zeky Murra-Anton, ISO New England, USA
zmurraanton@iso-ne.com

Roberto Serrano, Brown University, USA
roberto_serrano@brown.edu

ABSTRACT

Can private information or mediation change a sender's behavior and improve the receiver's expected utility in persuasive communication games? In a mediated Bayesian persuasion model, private information cannot improve the receiver's expected utility when the sender communicates it. When the intermediary communicates the private information, the receiver's expected utility improves only under sufficient accuracy of the intermediary's private information, as captured by a positive *autarky value of the intermediary's private information* (AVIPI). Finally, different classes of equilibria are analyzed to show that the sender's strategic behavior is generally affected by the intermediary's presence as he tries to persuade the intermediary to, in turn, persuade the receiver.

Keywords: Communication, intermediation, Bayesian persuasion.

JEL Classification Numbers: D82, D83, C72.

This paper was written for the special issue of this journal in honor of Professor Vincent Crawford's 75 birthday, which was celebrated at the 2024 Conference on Mechanism and Institution Design, Budapest, July 8-12, 2024. The paper has not been funded or endorsed by ISO New England; all views herein expressed are the authors' only. We thank Antonio Cabrales, Vincent Crawford, Philippe Jehiel, Emir Kamenica, Sam Kapon, Teddy Mekonnen, Giacomo Rubbini, Larry Samuelson, and Jesse Shapiro for their helpful comments. Serrano thanks Universidad Carlos III in Madrid for its warm hospitality and Fundacion ONCE for research support. We thank the editor and two anonymous referees who provided useful feedback that improved the paper.

1. INTRODUCTION

Consider a Ph.D. advisor who must write a recommendation letter to persuade a college to hire his student, who can be either a good or a bad teacher. The college, being uncertain of the student's true quality, typically appoints a hiring committee to evaluate the candidate and make a final recommendation independently. In this process, the sender attempts to persuade the college but communicates with it only through the hiring committee. Meanwhile, the hiring committee receives the advisor's letter and conducts additional evaluations based on independent sources. After considering both aspects, the committee provides a positive or negative hiring recommendation to the college, typically with only partial or no specific details about the outcome of the different aspects of the evaluation within the committee.

This scenario illustrates a Bayesian persuasion problem: an agent, the sender (he), strategically provides information to another agent, the receiver (she), to persuade her to take an action that may not align with her prior beliefs. However, unlike standard persuasion models, this setting involves two distinctive features that extend beyond traditional information design: (i) a communication intermediary (she) representing the receiver's interests, and (ii) a source of (possibly imperfect) objective information that the receiver learns from other agents. Notably, such features are present in various other applications, such as pharmaceutical research, corporate promotions, asset management, policy lobbying, and e-Commerce.[1]

This paper analyzes a stylized model of mediated persuasive communication where the sender and the intermediary can access private information that is unverifiable to the receiver. In standard persuasion models, absent an intermediary or private information, the sender can manipulate the receiver's beliefs by strategically obfuscating information, effectively extracting all surplus from the receiver (see Kamenica & Gentzkow (2011), Bergemann & Morris (2019)). In such cases, the receiver gains no additional utility from communication, as she is left indifferent between relying on her prior beliefs or following the sender's recommendation. This raises several key questions. Can an intermediary or private information improve the receiver's expected utility? Does the effectiveness of private information depend on who is conveying it? Can private information or an intermediary alter the sender's

[1] In Appendix B, we provide details about how our model can fit these applications.

communication strategy? This paper provides answers to these questions.

To illustrate our analysis, consider again the advisor-college example. The advisor (sender) aims to convince the college (receiver) to hire his student. To do this, the advisor writes a letter emphasizing select aspects of the student's qualifications. The college seeks to hire the student only if she is a good teacher, whereas the advisor's goal may be to get the student hired regardless of her teaching abilities. Due to information frictions, such as lack of expertise, the college appoints a hiring committee (intermediary) to recommend whether to hire the student. Finally, the committee, with the same preferences as the college, reads the advisor's letter and, along with independent sources of information such as the student's dossier, passes a final recommendation.

Theorem 1 establishes that, regardless of the evidence in the student's dossier, without the hiring committee, the advisor persuades the college to hire the student while leaving the college indifferent between hiring and not hiring. In other words, private information does not benefit the receiver when it is directly communicated by the sender. The reason is that the advisor can always manipulate the message in his letter by strategically obfuscating key details in the student's dossier. As a result, the advisor mimics the Bayesian persuasion communication strategy without private information: with probability one, the advisor truthfully reveals when the student is good, but with positive probability, he engages in *fake news* (misrepresenting a bad teacher as good).

Our main result, Theorem 2, characterizes the expected equilibrium utility of the receiver as a function of the precision of the private information that the mediator privately investigates. We introduce the *Autarky Value of the Intermediary's Private Information* (AVIPI). The AVIPI measures the maximum equilibrium expected utility gain the intermediary can secure for the receiver over allowing the sender to communicate directly, on the basis of her private information alone. In our example, the AVIPI is the level of utility the hiring committee secures for the college by only relying on its findings while looking at the student's dossier, but not the advisor's recommendation letter.

Formally, Theorem 2 establishes that the expected equilibrium utility of the receiver always is equal to the AVIPI. A strictly positive AVIPI empowers the hiring committee with two key tools to improve the utility of the college through direct communication: (i) the hiring committee can design a mechanism that selectively incorporates the advisor's recommendation when it benefits the receiver, and (ii) the hiring committee can completely disregard the advisor's communication when doing so increases the expected utility of the college. In equilibrium, this forces the advisor to tailor his communication strategy to ensure that part of his message is relayed to the college.

Notably, while the receiver's equilibrium expected utility equals the AVIPI, it does not necessarily imply that the intermediary always excludes the sender's recommendation in equilibrium. We demonstrate that when the intermediary considers the sender's recommendation but the receiver's expected utility exceeds the AVIPI, the sender can always devise an experiment that extracts all the receiver's additional surplus beyond the AVIPI. Consequently, in a model with obligatory mediation using an unbiased intermediary, the receiver avoids complete surplus extraction, attaining the same payoff as if the sender were absent, providing a robustness check on standard Bayesian persuasion techniques.

Our complementary results establish several equilibrium behaviors. In Proposition 2, we show that regardless of the accuracy of the student's dossier, there is always an equilibrium where the hiring committee ignores the sender's recommendation letter. However, under general conditions, there are no equilibria in which the hiring committee always recommends hiring the student, as they create opportunities for the advisor to manipulate the college decision. Similarly, Proposition 3 establishes that equilibria where the hiring committee relays the advisor's recommendation intact do not exist (unless the advisor insists on writing his "direct communication" letter and the dossier contains no information whatsoever). In addition, Proposition 4 rules out equilibria in which the hiring committee perfectly combines her findings in the student's dossier with the advisor's recommendation. Finally, Proposition 5 shows that, except in extreme cases where private information is either uninformative or perfectly identifies the good state, there is no equilibrium where the advisor's strategy remains unchanged by the presence of the hiring committee.

Our findings have important policy implications for mitigating misinformation. A relevant application is the moderation of social media content. For example, Twitter (now 'X') used to employ fact checking, content labeling, and *pre-bunking* to combat misinformation.[2] First, our results highlight the importance of independent communication of private information that aligns with the receiver's interests. Without independent verification, persuasive agents can distort facts, undermining the receiver's welfare. Second, our analysis underscores the need for accurate private information. Even with in-

[2] Pre-bunking refers to proactively countering misleading narratives through fact-checking, logic-checking, or source-checking.

dependent counter-messaging, ineffective private information fails to protect the receiver from persuasion if it lacks sufficient informativeness in the AVIPI sense. Without precise private information, the sender anticipates that the intermediary cannot effectively filter persuasive communication, allowing the sender to extract the receiver's surplus through misinformation.

We structure the remainder of the paper as follows. Section 2 presents the model. Section 3 provides benchmark results when the sender communicates private information directly to the receiver. Section 4 introduces our equilibrium model and constructs the AVIPI. Section 5 states our main findings. The related literature and concluding remarks are presented in Section 6 and Section 7, respectively. Proofs are relegated to the Appendix, which also contains a part on details of the applications.

2. MODEL

2.1. Actions and Payoffs

We consider the primitives of Bergemann & Morris's (2019) two-state binary action communication model. The state of the world can be good or bad, denoted by $t \in T \equiv \{t_g, t_b\}$. A sender (he), a receiver (she), and an intermediary (she), who are uninformed about the true state of the world, share a common uniform prior. The receiver must choose an action $a \in A \equiv \{a_1, a_2\}$; we sometimes refer to a_2 as *default action*. In Appendix B, we discuss how our model can be used in various practical problems.

We assume that the intermediary and the receiver share the same preferences. Table 1 summarizes the receiver/intermediary's (first entry) and sender's (second entry) payoffs from each state-action combination:

Table 1: Receiver's and Sender's Ex-post Payoffs.

	t_g	t_b
a_1	$x \in (0,1), 1$	$-1, 1$
a_2	$0, 0$	$0, 0$

Our framework is standard in information design:[3] the sender has state-

[3] Our model belongs to the class studied by Gentzkow & Kamenica (2016b) if we represent state

independent preferences and, under the prior, the receiver's default action a_2 is the sender's least preferred one.

2.2. Recommendation Experiments

We study Bayesian persuasion communication, where players choose and commit to experiments before any information related to the state of the world is realized. We focus on *recommendation experiments*, namely, experiments where the recommendation space coincides with the action space. In this section, we derive the obedience constraints and expected utility functions that we use in the rest of our analysis as a function of arbitrary experiments. We introduce the details of the experiments in the context of the sender-intermediary strategic interaction. For now, we denote the sender's experiment by σ and the intermediary's experiment by μ.

Consider an arbitrary recommendation experiment $e \in \{\sigma, \mu\}$. We denote by e_t the probability of experiment e recommending a_1 when the state of the world is t.[4] We say that an experiment is *obedient* when it is optimal for the receiver to follow its recommendation. The receiver follows a recommendation a_1 and a_2 only when, respectively,[5]

$$e_g x - e_b \geq 0. \tag{OC}$$

$$(1 - e_g)x - (1 - e_b) \leq 0. \tag{1}$$

As $x < 1$, Equation 1 is redundant to Equation OC, which we call *the obedience constraint*. When Equation OC is satisfied, the probability that the receiver chooses a_1 after a recommendation is the probability that the experiment recommends a_1. Thus, the expected utility of the sender is the total probability of recommending a_1:

$$U(e) = \frac{1}{2}(e_g + e_b). \tag{2}$$

Meanwhile, the expected utility of the receiver is

$$V(e) = \frac{1}{2}(xe_g - e_b). \tag{3}$$

t_g by $\omega = 1$, state t_b by $\omega = 0$, and write the receiver's preferences as $v(\omega) = \omega x + (1 - \omega)(-1)$.

4 For instance, σ_t is the probability of σ recommending a_1 in state t.

5 The receiver's posterior belief of the state t following a recommendation a_1 is $\frac{e_t}{e_t + e_{-t}}$.

Note that Equation OC is equivalent to $V(e) \geq 0$, which we use when suitable. In other words, the receiver obeys the experiment e only when the expected utility of doing so is greater than the utility of the default action.

The expected utility of the sender and the receiver shows a partial incentive misalignment. Both agents benefit from a higher probability of recommending a_1 when the state is t_g, whereas the sender benefits and the receiver suffers from a higher probability of recommending a_1 when the state is t_b. Finally, the maximum possible expected utility the sender and receiver can, respectively, attain under Equation OC are:

$$U^{\max} = \frac{1+x}{2}, \tag{4}$$

and

$$V^{\max} = \frac{x}{2}. \tag{5}$$

2.3. Private Information

In our model, there is a source of *private information* s about the state of the world, interpreted as a noisy signal about the true state. Formally, $s : T \to \triangle(T)$, with $s(t' \mid t)$ representing the probability that the realized signal is t' when the true state is t. We denote by $\hat{s}$ the realization of the signal s.

The *accuracy* of the private information is the probability that the signal reveals the true state of the world, namely $\epsilon_t \equiv s(\hat{s} = t \mid t)$. Private information is a *perfectly accurate signal* when $\epsilon_g = \epsilon_b = 1$ or when $\epsilon_g = \epsilon_b = 0$, as any realization $\hat{s}$ perfectly reveals the state of the world. Without loss of generality, we assume that $\epsilon_g + \epsilon_b \geq 1$.[6] We define the *combined informativeness of signal s* as the departure of s from a completely uninformative signal, namely $I(\epsilon_g, \epsilon_b) \equiv \epsilon_g + \epsilon_b - 1$.

2.4. Bayesian Persuasion Case

To add perspective to our results, we often pivot to the canonical case where the intermediary does not exist and the sender does not have private information

[6] Our model is completely symmetric when $\epsilon_g + \epsilon_b < 1$, as we can relabel signals so that the probabilities of the relabeled signals are $\epsilon'_g = 1 - \epsilon_g$ and $\epsilon'_b = 1 - \epsilon_b$. Perhaps an overkill in writing these proofs, but in an earlier version of the paper, we show that our results extend unaffected to $\epsilon_g + \epsilon_b < 1$.

(Kamenica & Gentzkow, 2011). We call this the Bayesian persuasion (BP) case. Formally, the sender-receiver interaction is as follows:

1. Before knowing the state of the world, the sender chooses and commits to an experiment $\sigma : T \to \triangle(A)$.

2. The state of the world is privately realized, and the sender recommends an action to the receiver according to σ.

3. The receiver, who knows the sender's experiment but not the state of the world, observes the sender's recommendation and updates her beliefs to make a decision.

In the BP case, the sender's problem is

$$\max_{\sigma} \frac{1}{2}(\sigma_g + \sigma_b) \text{ s.t. } \textit{Equation OC} \tag{BP}$$

The unique solution to the problem in Equation BP is to recommend a_1 with probability one when the state is t_g, and with probability x when the state is t_b (Bergemann & Morris, 2019). By recommending a_1 in state t_g, the sender maximizes the probability that the receiver is willing to tolerate, while still satisfying Equation OC, for recommending a_1 in state t_b. As a result, the sender leaves the receiver indifferent between following a recommendation and choosing the default action: the sender's and receiver's expected utilities are $U^{BP} = U^{max}$ and $V^{BP} = 0$. In the sequel, we call this solution the *Bayesian Persuasion Communication Policy (BPCP)*, denoted by σ^{BP}, with $\sigma_g^{BP} = 1$ and $\sigma_b^{BP} = x$.

3. DIRECT COMMUNICATION BENCHMARK

Our benchmark model is similar to the BP case, except that the sender has private information and communicates it directly to the receiver through his recommendation. Formally,

1. Before knowing the state of the world or the realization of his private information, the sender chooses and commits to an experiment $\sigma : T \times T \to \triangle(A)$.

2. The sender's private information and the state of the world are privately realized, and the sender recommends an action to the receiver according to σ.

3. The receiver, who knows the sender's experiment but not the state of the world or the sender's private information, observes the sender's recommendation and updates her beliefs to make a decision.

We denote by σ^{DC} the sender's experiment in the direct communication benchmark. The following theorem summarizes the sender's optimal communication strategy:

Theorem 1. *When the sender communicates directly to the receiver, any experiment σ^{DC} such that $\sum_{\hat{s}} \sigma^{DC}(a_1 \mid \hat{s}, t)s(\hat{s} \mid t) = \sigma_t^{BP}$ for all t is sender-optimal. Moreover, in any sender-optimal experiment, $U^{DC} = U^{max}$ and $V^{DC} = 0$.*

Theorem 1 establishes that under direct communication, the sender's optimal experiment mimics the BPCP. As a result, the sender extracts the receiver's whole expected utility, leaving her indifferent between following a recommendation and choosing the default action. Interestingly, the sender does so independently of the accuracy of the private information.

The result is a consequence of two facts. First, the expected utility of the sender and the obedience constraint depend only on the total probability of recommending a_1 in state t. Second, Lemma 1, stated below and key to the message of the theorem, establishes that the sender can *bend the facts*: he can manipulate his communication of the private information by suitably choosing $\sigma(a_1 \mid \hat{s}, t)$ to achieve any desired total probability of recommending a_1 in state t. As a result, the sender can target a probability $\sigma(a_1 \mid t)$ to solve the problem in Equation BP.

Lemma 1 (Private Information Suppression)**.** *For any state t and any $y \in [0, 1]$, there are probabilities $\sigma(a_1 \mid \hat{s}, t)$ such that*

$$y = \sum_{\hat{s}} \sigma(a_1 \mid \hat{s}, t)s(\hat{s} \mid t). \tag{IS}$$

One implication of Lemma 1 is that one can restrict attention to the sender's experiments where the signal is suppressed from the domain of the experiment, as he can always bend the facts and change the corresponding recommendation probabilities. We shall do so in the main model below.

4. EQUILIBRIUM INTERMEDIATION MODEL

4.1. Setup

We now introduce the intermediary in the game. The strategic interaction we analyze is the following.

1. Before knowing the state of the world or the realization of any private information, the sender and intermediary choose and commit simultaneously to experiments $\sigma : T \to \triangle(A)$ and $\mu : T \times A \to \triangle(A)$, respectively.

2. The state of the world is privately realized. Only the sender observes it and makes a recommendation to the intermediary according to σ.

3. The intermediary's private information is realized. Based on her private information and the sender's recommendation, the intermediary recommends an action to the receiver according to μ.

4. The receiver observes the intermediary's recommendation but not the sender's recommendation, the intermediary's private information, or the state of the world. Along with her knowledge of the sender's and intermediary's experiments, she updates her beliefs to make a decision.

Notice that communication from the sender to the receiver occurs exclusively through the intermediary. Two remarks regarding our model are on point. First, in the sender-intermediary game described above, the sender does not observe private information. As advanced at the end of the previous section, this simplification is without loss of generality due to Lemma 1: the sender's equilibrium payoff only depends on his total probability of recommending a_1, so, in equilibrium, his private information is irrelevant, as he mimics the case without private information.

Second, our model is strategically equivalent to one without an intermediary, where the receiver ex ante commits to a stochastic decision rule mapping the sender's realized message and the private information realization into actions. For the applications we have in mind, it is more natural to envision an intermediary instead of a receiver's stochastic plan.[7]

[7] In Section 6, we provide further detail and differences with the existing literature on information design with the receiver's commitment power.

We note that two assumptions are embedded in the sender-intermediary game described above. First, the receiver knows the distribution of the private information, but does not observe its realizations. Second, conditional on the state of the world, the intermediary's private information realization is independent of the sender's recommendation. Letting $\hat{s}$, $\hat{\sigma}$, and $\hat{\mu}$ be realizations of s, σ, and μ respectively, these two assumptions imply that the total probabilities that the sender and intermediary recommend $\hat{\sigma}$ and $\hat{\mu}$ in state t are

$$\sigma(\hat{\sigma} \mid t) \equiv \mathbb{P}(\hat{\sigma} \mid t). \tag{6}$$

and

$$\mu(\hat{\mu} \mid t) = \sum_{\hat{s},\hat{\sigma}} \mu(\hat{\mu} \mid \hat{s}, \hat{\sigma})\sigma(\hat{\sigma} \mid t)s(\hat{s} \mid t). \tag{7}$$

4.2. The Autarky Value of the Intermediary's Private Information

We start our analysis by noting that the intermediary has two strategies that provide a lower bound to the receiver's expected utility, independently of the sender's experiment: the intermediary can fully trust her private information or can always recommend a_2. When the intermediary uses either of these strategies, we say *the intermediary acts in autarky*.

First, suppose that the intermediary fully trusts her private information. Upon realizing $\hat{s} = t_g$, the intermediary acts as if she were convinced that the true state of the world is t_g, recommending a_1; similarly, when $\hat{s} = t_b$, her recommendation is a_2. Mathematically, $\mu(a_1 \mid t_g, \hat{\sigma}) = 1$ and $\mu(a_1 \mid t_b, \hat{\sigma}) = 0$ for all $\hat{\sigma}$. The receiver then knows that the total probability of receiving a recommendation a_1 when the state is t_g is $\mu_g = s(t_g \mid t_g) = \epsilon_g$ and when the state is bad is $\mu_b = s(t_b \mid t_g) = (1 - \epsilon_b)$. Hence, Equation 3 implies that the expected utility of the receiver is

$$\underline{V}(\epsilon_g, \epsilon_b) = \frac{1}{2}(x\epsilon_g - (1 - \epsilon_b)). \tag{8}$$

Second, the expected utility of the receiver is zero if the intermediary recommends a_2 with probability one. By relying on either of these strategies, the intermediary can guarantee a payoff for the receiver of $\alpha(\epsilon_g, \epsilon_b) =$

$\max\{0, \underline{V}(\epsilon_g, \epsilon_b)\} \geq 0$.[8] Note that $\alpha(\epsilon_g, \epsilon_b)$ depends on the accuracy of the private information of the intermediary. We refer to $\alpha(\epsilon_g, \epsilon_b)$ as the *Autarky Value of the Intermediary's Private Information (AVIPI)*, namely, the payoff the intermediary can secure for the receiver as a function of her private information when acting in autarky.

4.3. Equilibrium

More generally, the intermediary does not necessarily have to act in autarky. Next, we look for recommendation experiments μ^* and σ^* that constitute a Perfect Bayesian Equilibrium of the communication game.[9] An important challenge is to guarantee that the equilibrium experiments are obedient. A key simplifying observation is that the intermediary recommending the default action with probability one is always obedient, and so we can analyze the game where the obedience constraint in Equation OC is not actively considered.[10] As a result, any intermediary's best response to the sender's experiment cannot do worse than recommending the default action and will also be obedient. The following lemma formalizes this logic:

Lemma 2. *The best response of the intermediary to any σ is always obedient.*

Next, Proposition 1 establishes that an equilibrium exists for any parameters of the model.

Proposition 1. *For any $(\epsilon_g, \epsilon_b) \in [0, 1]^2$ such that $\epsilon_g + \epsilon_b \geq 1$ and $x \in (0, 1)$, an equilibrium (μ^*, σ^*) exists.*

8 Due to our model's symmetry, when $\epsilon_g + \epsilon_b < 1$, the intermediary replaces the "fully trusting her signal" strategy with "fully distrusting her signal," namely, recommending a_1 if and only if $\hat{s} = t_b$. In that case, $\underline{V}'(\epsilon_g, \epsilon_b) = \frac{1}{2}((1 - \epsilon_g)x - \epsilon_b)$ replaces $\underline{V}$ in the AVIPI. The reader must know that $\underline{V}'$ mechanically appears in some of our proofs. However, the term is merely a nuisance that eases the exposition, as $\epsilon_g + \epsilon_b \geq 1$ implies $\underline{V}' \leq 0$.

9 After incorporating the obedience constraint for the receiver, our equilibrium notion is a special case of the correlated Bayes-Nash equilibrium in Bergemann & Morris (2016). The two players aim to maximize expected payoffs with their designed obedient experiments, committing to the plan after every state or private information realization. No correlation device features in our model. Our notion of equilibrium is also a special case of the communication equilibrium in Forges (1986), reinterpreting the intermediary's experiment as a communication device that takes $(\hat{s}, \hat{\sigma})$ as inputs to produce $\hat{\mu}$ as output.

10 In other words, there is no equilibrium of the unrestricted game that is not an equilibrium of the game restricted by obedience.

Although this result gives us the existence of equilibrium, its uniqueness is not to be found. However, our results will characterize the relevant properties of all equilibria.

4.4. Best-Reply Functions and Equilibrium Expected Utility

It is helpful for our analysis to write the sender's and receiver's expected utility in terms of two objects of interest. The first is *the contribution to the receiver's expected utility from the intermediary recommending a_1 when her private information is $\hat{s}$ and the sender's recommendation is $\hat{\sigma}$*, denoted $\overline{v}_{s,\sigma}(\hat{s},\hat{\sigma})$. Denoting by $v(t)$ the receiver's payoff of choosing a_1 when the true state is t, then:[11]

$$\overline{v}_{s,\sigma}(\hat{s},\hat{\sigma}) \equiv \mathbb{P}(\hat{s},\hat{\sigma})\mathbb{E}(v(t) \mid \hat{s},\hat{\sigma}) = \frac{1}{2}\big(x\sigma(\hat{\sigma} \mid t_g)s(\hat{s} \mid t_g) - \sigma(\hat{\sigma} \mid t_b)s(\hat{s} \mid t_b)\big). \quad (9)$$

The second object is *the average increase in the probability that the intermediary recommends a_1, when the sender recommends a_1 instead of a_2 in state t*, denoted by $\overline{\triangle}_{\mu,s}(t)$. To formally define $\overline{\triangle}_{\mu,s}(t)$, we note that the change in the probability that the intermediary recommends a_1 when she has private information $\hat{s}$ and the sender changes his recommendation from a_2 to a_1 is $\triangle\mu(\hat{s}) = \mu(a_1 \mid \hat{s}, a_1) - \mu(a_1 \mid \hat{s}, a_2)$, so that

$$\overline{\triangle}_{\mu,s}(t) \equiv \mathbb{E}(\triangle\mu(\hat{s}) \mid t) = \sum_{\hat{s}} \triangle\mu(\hat{s})s(\hat{s} \mid t). \quad (10)$$

To avoid cumbersome notation, we suppress the dependence of $\overline{v}$ and $\overline{\triangle}$ on s, σ, and μ, except when necessary to avoid ambiguity. On the one hand, $\overline{v}$ determines how the sender's experiment affects the receiver's expected utility, so it determines the intermediary's best-response behavior to σ. On the other hand, $\overline{\triangle}$ determines how the intermediary's experiment changes when the sender changes her recommendation, so it determines the sender's best-response behavior to μ. The following lemma characterizes the intermediary's and sender's expected payoffs and best responses in terms of $\overline{v}$ and $\overline{\triangle}$:

---

[11] Mathematically, if $\mathbb{I}$ is the indicator function, $\overline{v}_{s,\sigma}(\hat{s},\hat{\sigma}) = \mathbb{E}(v(t)\mathbb{I}_{(\hat{s},\hat{\sigma})})$, interpreted as the contribution of the realization $(\hat{s},\hat{\sigma})$ to the unconditional (ex-ante) expected payoff $\mathbb{E}(v(t))$.

Lemma 3. *The intermediary's expected payoff and (unrestricted) best-response correspondence are, respectively,*

$$V(\mu,\sigma) = \sum_{\hat{\sigma},\hat{s}} \mu(a_1 \mid \hat{s},\hat{\sigma})\overline{v}(\hat{s},\hat{\sigma}), \tag{11}$$

and

$$\mu^*(a_1 \mid \hat{s},\hat{\sigma}) = \begin{cases} 1 & \text{if } \overline{v}(\hat{s},\hat{\sigma}) > 0, \\ \in [0,1] & \text{if } \overline{v}(\hat{s},\hat{\sigma}) = 0, \\ 0 & \text{if } \overline{v}(\hat{s},\hat{\sigma}) < 0. \end{cases} \tag{12}$$

The sender's expected payoff and (unrestricted) best-response correspondence are, respectively,

$$U(\mu,\sigma) = \frac{1}{2}\left(\sum_t \sigma(a_1 \mid t)\overline{\triangle}(t) + \sum_{t,\hat{s}} s(\hat{s} \mid t)\mu(a_1 \mid \hat{s}, a_2)\right), \tag{13}$$

and

$$\sigma^*(a_1 \mid t) = \begin{cases} 1 & \text{if } \overline{\triangle}(t) > 0, \\ \in [0,1] & \text{if } \overline{\triangle}(t) = 0, \\ 0 & \text{if } \overline{\triangle}(t) < 0. \end{cases} \tag{14}$$

Lemma 3 reveals key insights into the strategic interaction between the sender and the intermediary. On the one hand, the intermediary is only willing to recommend a_1 with positive probability when the contribution to the receiver's expected utility is strictly positive. In other words, the intermediary's optimal strategy is to recommend a_2 when the sender's recommendation and the private information suggest that recommending a_1 is expected to harm the receiver. On the other hand, the sender understands the optimal strategy of the intermediary, so he recommends a_1 with positive probability only if it is expected to increase the average probability of the intermediary recommending a_1.

5. MAIN RESULTS IN THE EQUILIBRIUM INTERMEDIATION MODEL

5.1. The Receiver's Equilibrium Welfare and the AVIPI

Given that the intermediary can always secure the AVIPI for the receiver, regardless of the sender, it is natural to investigate its relationship with the receiver's equilibrium expected utility. More precisely, can the sender still persuade the receiver through the intermediary in a way that extracts most or all of her surplus?

The following Theorem, our main result, answers these questions: the receiver's equilibrium expected utility is always equal to the AVIPI. Therefore, the intermediary can avoid the extraction of the whole surplus if and only if the AVIPI is strictly positive.[12]

Theorem 2. *If (μ^*, σ^*) is an equilibrium, then $V(\mu^*, \sigma^*) = \alpha(\epsilon_g, \epsilon_b)$. Moreover, the following statements are equivalent:*

(i) $V(\mu^*, \sigma^*) > 0$.

(ii) There is $(\hat{s}, \hat{\sigma})$ such that $\overline{v}(\hat{s}, \hat{\sigma}) > 0$.

The equivalence between (i) and (ii) in Theorem 2 follows mechanically from Lemma 3. Specifically, it is optimal for the intermediary to recommend a_1 if and only if her private information-recommendation pair $(\hat{s}, \hat{\sigma})$ positively contributes to the expected utility of the receiver, namely $\overline{v}(\hat{s}, \hat{\sigma}) > 0$. If $(\hat{s}, \hat{\sigma})$ contributes negatively, the intermediary can neutralize such a negative effect by "playing safe" and recommending a_2 with probability one.

Regarding the equilibrium receiver's payoff, the first thing to note is that when two experiments (μ, σ) combine into an expected utility of $V(\mu, \sigma) < \alpha(\epsilon_g, \epsilon_b)$, the intermediary can improve the receiver's outcome by pivoting back to acting in autarky, the case in which she secures a payoff of $\alpha(\epsilon_g, \epsilon_b) \geq 0$. Therefore, in any equilibrium, we must have $V(\mu^*, \sigma^*) \geq \alpha(\epsilon_g, \epsilon_b)$.

Then, given the equilibrium strategy of the intermediary μ^*, the sender's equilibrium strategy must maximize $U(\mu^*, \sigma)$ subject to $V(\mu^*, \sigma) \geq \alpha(\epsilon_g, \epsilon_b)$. The rest of the proof consists of showing that $V(\mu^*, \sigma^*) > \alpha(\epsilon_g, \epsilon_b)$ is impossible. Note that this is not immediate as in the BP case, because $V(\mu, \sigma)$ may not be strictly monotonic in σ in a way that unambiguously an increase in σ_b increases the sender's expected utility but makes Equation OC harder to satisfy at the same time, leading to the obedience constraint to bind.

[12] An earlier version of the paper also contained a different interaction between the sender and the intermediary. We referred to it as "surprise intervention,", which means that the sender did not strategize for the presence of the intermediary and always insisted on the BPCP. In that model, combined informativeness was the key statistic to transfer surplus back to the receiver, not the AVIPI.

Instead, the proof shows that if $V(\mu^*, \sigma) > \alpha(\epsilon_g, \epsilon_b)$, the intermediary must be accounting for part of the sender's recommendation in her experiment. Otherwise, she would trivially act in autarky, leading to $V(\mu^*, \sigma) = \alpha(\epsilon_g, \epsilon_b)$. However, knowing that the intermediary will rely on part of the sender's recommendation, he can find a deviation that takes advantage of such a communication channel with the receiver to strictly increase the probability that a_1 is recommended in equilibrium, a profitable deviation σ'. If such a profitable deviation is small enough, then we will have $V(\mu^*, \sigma') > \alpha(\epsilon_g, \epsilon_b)$, which means that the intermediary is still better off tolerating the deviation rather than acting in an autarky way. But then, of course, this logic can be repeated, leading to further increases in the sender's payoff at the expense of the intermediary's utility. However, when $V(\mu^*, \sigma) = \alpha(\epsilon_g, \epsilon_b)$, the sender's ability to profitably increase the probability that a_1 is recommended in equilibrium disappears. This is because the intermediary's best response is such that when the sender changes her strategy, the intermediary understands that $V(\mu^*, \sigma') < \alpha(\epsilon_g, \epsilon_b)$, a case in which she already planned to pivot to acting in autarky, an outcome that cannot strictly benefit the sender.

The following example shows the logic behind Theorem 2. Suppose $\epsilon_g = \frac{3}{4}$, $\epsilon_b = \frac{5}{8}$, and $x = \frac{5}{6}$. Then, $\underline{V}(\epsilon_g, \epsilon_b) = \frac{1}{8}$, and $\alpha(\epsilon_g, \epsilon_b) = \frac{1}{8}$.

First, assume that the intermediary is overly optimistic and expects the sender to benefit the receiver by designing an experiment $\sigma_g = 1$ and $\sigma_b = 0$. Then, the contribution to the receiver's expected utility of each private information-recommendation pair is

$$\overline{v}(\hat{s}, \hat{\sigma}) = \begin{cases} \frac{5}{16}, & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_1), \\ \frac{5}{48}, & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_1), \\ -\frac{3}{16}, & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_2), \\ -\frac{5}{16}, & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_2). \end{cases}$$

In this case, $\overline{v}(\hat{s}, a_1) > 0$ and $\overline{v}(\hat{s}, a_2) < 0$ for all $\hat{s}$, so the intermediary knows that passing the sender's recommendation intact to the receiver is beneficial and sets an optimal experiment according to Equation 12:

$$\mu^*(a_1 \mid \hat{s}, \hat{\sigma}) = \begin{cases} 1, & \text{if } \hat{\sigma} = a_1, \\ 0, & \text{if } \hat{\sigma} = a_2. \end{cases}$$

Then, the receiver's expected utility is $V(\mu^*, \sigma) = \frac{5}{12} > \alpha(\epsilon_g, \epsilon_b)$. However, σ is not optimal for the sender. Mathematically, the sender knows that the best response of the intermediary implies $\overline{\triangle}(t_b) = 1$, so his best response according to Equation 14 is $\sigma_b^* = 1 \neq \sigma_b$. We conclude that $\sigma_g = 1$ and $\sigma_b = 0$ cannot be part of an equilibrium. The intuition is that the sender understands that a recommendation a_1 will be passed intact to the receiver, so he can strictly benefit by increasing σ_b by a small amount, increasing the probability that a_1 is the recommendation for the receiver.

We show that this deviation channel disappears for the sender when $V(\mu, \sigma) = \alpha(\epsilon_g, \epsilon_b)$. To that end, suppose that the intermediary expects $\sigma_g^* = \sigma_b^* = \frac{1}{2}$. Then, the contribution to the receiver's expected utility of each private information-recommendation pair is

$$\overline{v}(\hat{s}, \hat{\sigma}) = \begin{cases} \frac{1}{16}, & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_1), \\ -\frac{5}{48}, & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_1), \\ \frac{1}{16}, & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_2), \\ -\frac{5}{48}, & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_2). \end{cases}$$

In this case, $\overline{v}(t_g, \hat{\sigma}) > 0$ and $\overline{v}(t_b, \hat{\sigma}) < 0$ for all $\hat{\sigma}$, so the intermediary knows that passing the sender's recommendation intact to the receiver is not beneficial and sets an optimal experiment according to Equation 12:

$$\mu^*(a_1 \mid \hat{s}, \hat{\sigma}) = \begin{cases} 1, & \text{if } \hat{s} = t_g, \\ 0, & \text{if } \hat{s} = t_b. \end{cases}$$

Then, the receiver's expected utility is $V(\mu^*, \sigma) = \frac{1}{8} = \alpha(\epsilon_g, \epsilon_b)$. In this case, $\overline{\triangle}(t) = 0$ for all t, so the sender is left indifferent between σ^* and any other deviation. Given the sender's original choice, the intermediary's best response is to listen only to $\hat{s}$, securing a payoff equal to the AVIPI for the receiver. This is because σ^* is such that if the intermediary listens to the sender, for instance, by setting $\mu(a_1 \mid t_b, a_1) > 0$, this harms the expected utility of the receiver, namely $\overline{v}(t_b, a_1) < 0$. The proof of Theorem 2 shows that this logic extends to all equilibria.

A critical remark is that Theorem 2 does not imply that the intermediary relies only on her private information in equilibrium, leading to the AVIPI. Instead, Theorem 2 implies that relying only on her private information is enough to incentivize the sender, in equilibrium, to use a communication policy that leaves the receiver with a reward equal to AVIPI. In Section 5.3, we explore the different types of equilibria that may arise in our communication game and provide an example that shows that the intermediary may rely on the sender's recommendation in non-trivial ways (Section 5.3.4).

5.2. The Relationship Between the Intermediary's Private Information and the Receiver's Equilibrium Welfare

Theorem 2 shows that the AVIPI is crucial in helping the receiver attain a payoff above the Bayesian persuasion payoff. Naturally, the AVIPI is related to the accuracy of s and, hence, to its total combined informativeness. In this section, we explore how much information is necessary and sufficient to improve the receiver's payoff in the BP case.

Corollary 1. *Suppose* (μ^*, σ^*) *is an equilibrium. Then, the following statements hold:*

(i) $I(\epsilon_g, \epsilon_b) > 1 - x$ *is sufficient (but not necessary) for* $V(\mu^*, \sigma^*) > 0$.

(ii) $I(\epsilon_g, \epsilon_b) > 0$ *is necessary (but not sufficient) for* $V(\mu^*, \sigma^*) > 0$.

Corollary 1 distills one of the main economic insights of Theorem 2: some private information ($I(\epsilon_g, \epsilon_b) > 0$) is necessary to secure a strictly positive payoff for the receiver, but it may not be enough. Instead, an intermediary requires enough private information, in the AVIPI sense. Figure 1 provides a graphical representation of Corollary 1 in the (ϵ_g, ϵ_b) Cartesian plane.

The red line connecting $(0, 1)$ to $(1, 1 - x)$ is the locus of accuracy levels such that $\underline{V}(\epsilon_g, \epsilon_b) = \frac{1}{2}(x\epsilon_g - (1 - \epsilon_b)) = 0$. The accuracy levels in the blue region above $\underline{V}(\epsilon_g, \epsilon_b) = 0$ guarantee a strictly positive AVIPI (R_1 and R_2 in the figure). The accuracy levels in the regions R_1 and R_2 can be interpreted as those that are *sufficiently accurate* to guarantee a strictly positive AVIPI. Meanwhile, the red region between the locus $\underline{V}(\epsilon_g, \epsilon_b) = 0$ and the no-information locus $I(\epsilon_g, \epsilon_b) = 0$ is where the intermediary has some information, but not enough to grant a positive AVIPI. This region is represented by R_3 in the figure.

The region R_2 shows that $I(\epsilon_g, \epsilon_b) > 1 - x$ is sufficient but unnecessary for a strictly positive AVIPI, namely, part (i) of Corollary 1. Meanwhile, the existence of R_3, bounded below by $I(\epsilon_g, \epsilon_b) = 0$, demonstrates that $I(\epsilon_g, \epsilon_b) > 0$ is necessary but insufficient for a strictly positive AVIPI, namely, part (ii) of Corollary 1.

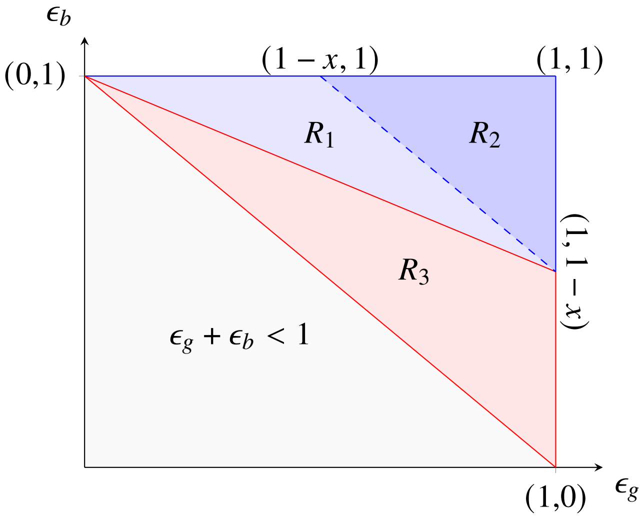

Figure 1: Informativeness Regions and Receiver's Equilibrium Expected Utility.

Only the accuracy levels in the blue regions provide enough leverage for the intermediary to help the receiver attain strictly positive equilibrium payoffs. Therefore, we refer to R_1 and R_2 as the *antipersuasive region*. Meanwhile, the accuracy levels in the red region, including the boundary, lead to a receiver's equilibrium payoff equal to the Bayesian persuasion case where the intermediary does not exist. Hence, we refer to R_3 as the *persuasive region*.

5.3. Classes of Equilibria

Our final goal is to understand how the sender and intermediary plan for each other in equilibrium. In this subsection, we analyze three natural classes of equilibria. When the intermediary ignores the sender, when she passes along the sender's recommendation, and when she perfectly blends her private information with the sender's recommendation. We conclude by analyzing whether the Bayesian Persuasion communication policy can be part of an equilibrium when there is an intermediary.

5.3.1. Ignoring the Sender's Recommendation

Theorem 2 relies on the fact that a positive AVIPI provides enough leverage to the intermediary to mitigate the persuasive efforts of the sender, securing a strictly positive payoff for the receiver. In this section, we study the conditions under which the intermediary can rely solely on her private information in equilibrium.

We say that *the intermediary ignores the sender's recommendation when her private information is* $\hat{s}$ when $\mu(a_1 \mid \hat{s}, a_1) = \mu(a_1 \mid \hat{s}, a_2) \equiv \mu(\hat{s})$. We interpret $\mu(\hat{s})$ as the probability of recommending a_1 after $\hat{s}$. If the intermediary ignores the sender's recommendation for all $\hat{s}$, we say that *the intermediary always ignores the sender's recommendation.* The following result establishes when the intermediary can ignore the sender in equilibrium:

Proposition 2. *The following statements are true:*

(i) There exists an equilibrium (μ^*, σ^*) *where the intermediary always ignores the sender's recommendation if and only if* $\overline{v}(\hat{s}, a_1)\overline{v}(\hat{s}, a_2) \geq 0$ *for all* $\hat{s}$.

(ii) For any accuracy levels, there is an equilibrium where the intermediary always ignores the sender's recommendation.

(iii) In an equilibrium where the intermediary always ignores the sender's recommendation, the only possibilities are the following:

$$\mu^*(\hat{s} = t_g) = \begin{cases} 0, & \textit{if } \underline{V} < 0, \\ 1, & \textit{if } \underline{V} > 0, \\ \in [0, 1], & \textit{if } \underline{V} = 0. \end{cases}$$

$$\mu^*(\hat{s} = t_b) = \begin{cases} 0, & \textit{if } \frac{1}{2}((1 - \epsilon_g)x - \epsilon_b) < 0, \\ \in [0, 1], & \textit{otherwise} . \end{cases}$$

Part (i) follows from the fact that if $\overline{v}(\hat{s}, a_1)\overline{v}(\hat{s}, a_2) < 0$, the intermediary would benefit from being responsive to the sender's recommendation. For example, if $\overline{v}(\hat{s}, a_1) > 0$ and $\overline{v}(\hat{s}, a_2) < 0$, then Equation 12 implies that the intermediary's best response is $\mu^*(a_1 \mid \hat{s}, a_1) = 1$ and $\mu^*(a_1 \mid \hat{s}, a_2) = 0$.

Part (ii) explores whether the necessary and sufficient condition established in part (i) imposes further restrictions on the accuracy levels that can support the ignoring of the sender in equilibrium. The answer is negative: such behavior can be sustained in equilibrium independently of the accuracy of the intermediary's private information. Trivially, $\sigma_g^* = \sigma_b^* = 0$ always satisfies the condition in part (i), regardless of (ϵ_g, ϵ_b).

Finally, part (iii) outlines how the intermediary shapes her optimal strategy as a function of the information she has. In the anti-persuasive region, $\underline{V}(\epsilon_g, \epsilon_b) > 0$, so the intermediary can comfortably trust a signal $\hat{s} = t_g$ and recommend a_1 with probability one, achieving a strictly positive payoff for the receiver. In the same region, a realization $\hat{s} = t_b$ is enough confirmation that the state of the world is likely to be t_b, which implies that $\frac{1}{2}((1-\epsilon_g)x - \epsilon_b) < 0$, so the intermediary is comfortable always recommending a_2.

Meanwhile, if the intermediary's private information is inaccurate enough, she understands that she is in the persuasive region where $\underline{V}(\epsilon_g, \epsilon_b) < 0$. Then, by realizing $\hat{s} = t_g$, the intermediary is not convinced enough that the state of the world is t_g and can avoid harm to the receiver by simply recommending a_2 with probability one. In this region, we have that $\frac{1}{2}((1 - \epsilon_g)x - \epsilon_b) \leq 0$ as well, so the intermediary also distrusts a realization $\hat{s} = t_b$, leading to a recommendation of a_2 with probability one.

5.3.2. Passing the Sender's Message Along

In this section, we study whether, in equilibrium, the threat of ignoring the sender provides enough incentives for him to choose an experiment producing recommendations that the intermediary optimally passes intact to the receiver.

We say that *the intermediary passes the sender's message along intact* when she perfectly mimics the sender's recommendation, namely $\mu(a_1 \mid \hat{s}, a_1) = 1$ and $\mu(a_1 \mid \hat{s}, a_2) = 0$ for all $\hat{s}$. The following proposition determines whether passing the sender's message along is possible in equilibrium:

Proposition 3. *There exists an equilibrium (μ^*, σ^*) where the intermediary always passes the sender's message along if and only if the intermediary has no combined informativeness and the sender uses the BPCP, i.e. $I(\epsilon_g, \epsilon_b) = 0$ and $\sigma_g^* = 1$, $\sigma_b^* = x$.*

Proposition 3 establishes that the intermediary always passes the sender's message in equilibrium only under very particular conditions. To understand the result, note that when the intermediary passes the sender's message along, the problem is equivalent to one where the intermediary is absent. From the sender's perspective, the only solution to this problem is the BPCP, which yields an expected utility of zero for the receiver.

The only case where the intermediary's optimal strategy is passing the message along when the sender uses the BPCP is when $I(\epsilon_g, \epsilon_b) = 0$, because she does not have any information and, hence, she only learns about the state of the world through the sender. However, when $I(\epsilon_g, \epsilon_b) > 0$, the intermediary has some confirmation about the state of the world, independent of the sender. In this case, the intermediary realizes that she can do better by contradicting the sender in some cases. For example, if $I(\epsilon_g, \epsilon_b) > 0$, $\sigma_g^* = 1$, and $\sigma_b^* = x$, $\overline{v}(t_b, a_1) = x(1 - \epsilon_g - \epsilon_b) < 0$, and so the intermediary prefers to deviate from passing the message along to $\mu(a_1 \mid t_b, a_1) = 0$. Section 5.3.5 below details the analysis of such a case.

5.3.3. Perfectly Blending Private Information with the Sender's Recommendation

Thus far, we have studied two extreme classes of equilibria. In Section 5.3.1, we investigated the case where the intermediary ignores the sender. Meanwhile, in Section 5.3.2, we studied the case in which the intermediary listens perfectly to the sender. In this section, we investigate an intermediate case, namely, where the intermediary relies both on her private information and the sender's recommendation.

We say that *the intermediary perfectly blends her private information with the sender's recommendation* when $\mu(a_1 \mid \hat{s}, \hat{\sigma}) \in (0, 1)$ for all $(\hat{s}, \hat{\sigma})$. Note that from Equation 9, we can compute $\overline{v}(t_g, a_2) = \underline{V}(\epsilon_g, \epsilon_b) - \overline{v}(t_g, a_1)$ and $\overline{v}(t_b, a_2) = \frac{1}{2}((1 - \epsilon_g)x - \epsilon_b) - \overline{v}(t_b, a_1)$. According to Equation 12, the intermediary only agrees to perfectly blend her private information with the sender's recommendation in equilibrium if $\overline{v}(\hat{s}, \hat{\sigma}) = 0$ for all $(\hat{s}, \hat{\sigma})$, which only happens if $\underline{V}(\epsilon_g, \epsilon_b) = \frac{1}{2}((1 - \epsilon_g)x - \epsilon_b) = 0$. However, the latter is not possible, as it requires that $x = 1$, whereas the model assumption is $x < 1$. We formalize this reasoning in the following result.

Proposition 4. *There is no equilibrium where the intermediary perfectly blends her private information with the sender's recommendation.*

The intuition behind Proposition 4 comes from the indifference that the intermediary can sustain. According to Equation 12, $\mu(a_1 \mid \hat{s}, \hat{\sigma}) \in (0, 1)$ for all $(\hat{s}, \hat{\sigma})$ is only possible when the intermediary is exactly indifferent to the contribution to the expected utility of every possible realization $(\hat{s}, \hat{\sigma})$, namely $\overline{v}(\hat{s}, \hat{\sigma}) = 0$ for all $(\hat{s}, \hat{\sigma})$. However, this is only possible if the reward for correctly recommending a_1 (x) and the penalty for incorrectly recommending a_1 (-1) are symmetric, hence the requirement of $x = 1$ for this class of equilibria. In our model, the penalty exceeds the reward and this indifference cannot be sustained.

5.3.4. *Nontrivial Equilibria*

Proposition 2, Proposition 3, and Proposition 4 may seem to suggest that the intermediary secures the AVIPI for the receiver in equilibrium in trivial ways, namely by ignoring the sender or by not having any information. However, this is not the case. We provide an example that shows how an informed intermediary can listen to the sender in relevant cases.

Suppose that $\sigma_1^* = 1$ and $\sigma_b^* = x = \frac{4}{5}$. Moreover, assume that $\epsilon_g = \frac{3}{4}$ and $\epsilon_b = 1$, so $\alpha(\epsilon_g, \epsilon_b) = \frac{3}{10}$. In this example, the intermediary has enough private information to guarantee a strictly positive AVIPI. Then, the contribution to the receiver's expected utility of each private information-recommendation pair is

$$\overline{v}(\hat{s}, \hat{\sigma}) = \begin{cases} \frac{3}{10}, & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_1), \\ -\frac{3}{10}, & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_1), \\ 0, & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_2), \\ -\frac{1}{10}, & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_2). \end{cases}$$

This implies via Equation 12, that the optimal intermediary policy is $\mu^*(a_1 \mid t_g, a_1) = 1$, $\mu^*(a_1 \mid t_b, a_1) = \mu^*(a_1 \mid t_b, a_2) = 0$, and $\mu^*(a_1 \mid t_g, a_2) \in [0, 1]$. In particular, let $\mu^*(a_1 \mid t_g, a_2) = 0$. Finally, we verify that σ^* is optimal given μ^*. In this case, we have $\triangle\mu(t_g) = 1$ and $\triangle\mu(t_b) = 0$, leading to $\overline{\triangle}(t_g) > 0$ and $\overline{\triangle}(t_b) = 0$. As a result, Equation 14 implies that $\sigma_g^* = 1$ and $\sigma_b^* = \frac{4}{5}$ can be sustained, confirming that (μ^*, σ^*) is an equilibrium. We have found an equilibrium where the intermediary follows the sender's recommendation when $\hat{s} = t_g$ and ignores the sender's recommendation when $\hat{s} = t_b$. It can be shown that a different possibility is when $\mu^*(a_1 \mid t_g, a_2) = \frac{1}{2}$, all else equal, a case in which the intermediary blends part of the sender's recommendation when $\hat{s} = t_g$.

5.3.5. Bayesian Persuasion Revisited

The previous example and Proposition 3 show that the BPCP may arise in equilibrium. The following result investigates under which circumstances the BPCP can be sustained as part of an equilibrium.

Proposition 5. *The BPCP features in equilibrium only in the following cases:*

(i) If the intermediary's private information is combined-uninformative, namely, $I(\epsilon_g, \epsilon_b) = 0$. Then $U(\mu^*, \sigma^*) = U^{BP} = \frac{1+x}{2}$ and $V(\mu^*, \sigma^*) = 0$.

(ii) The intermediary's private information is combined-informative, namely $I(\epsilon_g, \epsilon_b) > 0$ and the intermediary is perfectly informed about the good state of the world ($\epsilon_b = 1$ and $\epsilon_g > 0$). Moreover, in any such equilibria $U(\mu^*, \sigma^*) = \frac{\epsilon_g}{2}$ and $V(\mu^*, \sigma^*) = \epsilon_g V^{max} = \frac{\epsilon_g x}{2}$.

Proposition 5 confirms the findings in Proposition 3, that is, that the BPCP is an equilibrium possibility only under very particular conditions. Part (i), which follows immediately from Proposition 3, is included in the result for completeness and as an essential point.

Part (ii) provides further insight into how the intermediary protects the receiver against persuasive communication when the sender is set to use the BPCP. Note that the intermediary agrees with the sender's BPCP when the state is t_g: as $\sigma_g^* = 1$, the sender recommends a_1 only when the state is t_g, which fully benefits the receiver. However, the intermediary disagrees with the sender's BPCP when the state is t_b: as $\sigma_b^* = x$, the sender sometimes recommends a_1 when the state is t_b.

In general, without further information, the latter case makes it impossible for the intermediary to confirm that the state is t_g when $\hat{\sigma} = a_1$. As argued after Proposition 3, the intermediary can do better by shutting down some of the sender's recommendations under BPCP, incentivizing the sender to adopt a different strategy. The proof of Proposition 5 shows that this logic extends beyond $I(\epsilon_g, \epsilon_b) = 0$, except in the case where the intermediary can perfectly confirm that the state of the world is t_g. This is the case where $\epsilon_b = 1$ and $\epsilon_g > 0$, leading to $\mathbb{P}(t_g \mid \hat{s} = t_g) = \frac{\epsilon_g}{\epsilon_g + (1-\epsilon_b)} = 1$.

When the intermediary is perfectly informed about the good state of the world and the sender uses the BPCP, the intermediary's best response is to recommend a_1 only when she can confirm that the state of the world is t_g, which happens with ex-ante probability $\frac{\epsilon_g}{2}$. As a result, when recommending

a_1, the receiver is perfectly informed about the state of the world and attains her maximum possible payoff in Equation 5, namely x with probability $\frac{\epsilon_g}{2}$. The sender keeps the rest of the Bayesian Persuasion surplus in Equation 4, namely, a payoff of 1 with the same ex ante probability $\frac{\epsilon_g}{2}$. Meanwhile, when the intermediary recommends a_2, both the sender and the receiver receive a zero payoff. Hence, the sender's expected utility is $\frac{\epsilon_g}{2}$ and the receiver's expected utility is $\frac{\epsilon_g x}{2}$.

Part (ii) of Proposition 5 reveals two insights on the role of the intermediary. First, the intermediary can recover part of the receiver's surplus even when the sender uses the BPCP. Second, social welfare, along with the payoffs of the sender and the receiver, is strictly increasing in the intermediary's private information, showcasing a non-trivial role for the intermediary.

In such a case, the social surplus is $\frac{\epsilon_g(x+1)}{2}$, that is, a weighted average between confirming t_g and creating the Bayesian Persuasion surplus of $U^{BP} = \frac{x+1}{2}$ and not confirming t_g and recommending a_2 with probability one, which is zero. Note that the social surplus is strictly increasing in the intermediary's private information (ϵ_g), as are the receiver's and sender's expected utility. Note also that a big discontinuity of total equilibrium surplus at the point $(\epsilon_g, \epsilon_b) = (0, 1)$ between parts (i) and (ii) of the proposition, is caused by the starkly different behavior of the intermediary, i.e., from passing the sender's recommendation intact in part (i) to recommending a_1 only when the good signal is received in part (ii).

Finally, Figure 2 shows the levels of accuracy that support the BPCP in equilibrium. The figure presents the same accuracy regions in Figure 1. The black counter-diagonal depicts part (i) of Proposition 5, where $I(\epsilon_g, \epsilon_b) = 0$. Meanwhile, the horizontal line connecting $(0, 1)$ and $(1, 1)$ represents part (ii).

6. RELATED LITERATURE

We extend Kamenica & Gentzkow's (2011) Bayesian persuasion framework to accommodate a communication intermediary and private information. Conceptually, the two papers closest to our work are Kolotilin et al. (2017) and Perez-Richet & Skreta (2022). Kolotilin et al. (2017) studies persuasion mechanisms that condition the sender's experiment on a receiver's report of her private information. Our work differs mainly in two ways. First, the private information that our intermediary possesses is state-relevant, not payoff-

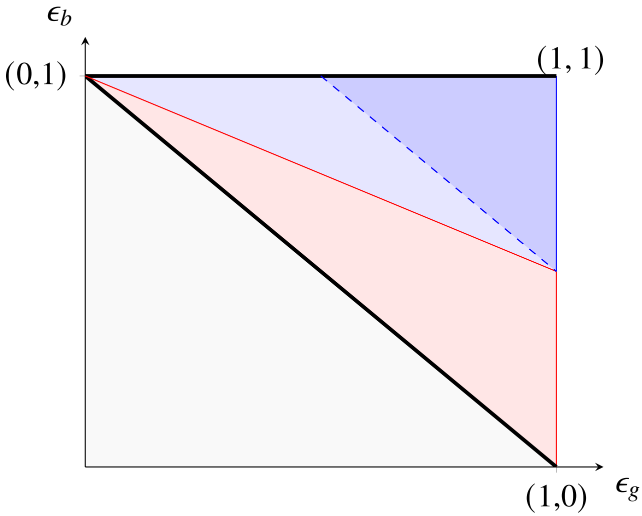

Figure 2: Locus of Accuracy Levels Supporting the BPCP as an Equilibrium.

relevant.[13] Second, consistent with our applications, we do not allow agents to communicate before the information structures are considered. Perez-Richet & Skreta (2022) studies optimal test design when the sender can falsify the state of the world after paying a cost. The main difference from our work is the existence of evidence available to both the sender and intermediary.

Another closely related paper is Bergemann & Morris (2019), since we borrowed the primitives of their model. Beyond the model, our paper has several differences from Bergemann & Morris (2019). First, we study the case of multiple informed senders with an uninformed receiver, while Bergemann & Morris (2019) analyzes the case where receivers are privately informed. Second, Bergemann & Morris (2019) focuses on the effect of the distinction between public or private messages on equilibrium behavior. In contrast, we focus on the impact of the identity of the agent communicating the private information. Finally, while both papers analyze the effect of the amount of information on equilibrium outcomes, Bergemann & Morris (2019) quantifies the amount of information in terms of the Blackwell order, while our analysis identifies a different metric of information, namely the AVIPI.

Our work also relates to the literature on information design with media-

[13] Equivalence is not guaranteed since Kolotilin et al. (2017) requires that the receiver's payoff is linear in the parameter over which she possesses private information.

tors. Ivanov (2010) studies whether a principal can improve direct communication by choosing among multiple uninformed and self-interested mediators. Kosenko (2020) studies equilibrium information revelation under an uninformed, self-interested mediator. In contrast, we focus on the role of a privately-informed mediator representing the receiver's interests in mitigating persuasive communication and on how her efficacy depends on the quality of her private information.

Several articles study sequential mediation by an arbitrary number of mediators. Ambrus et al. (2013), extends Crawford & Sobel's (1982) cheap-talk framework to study whether mediated communication improves outcomes over direct communication. In the problem of hierarchical Bayesian persuasion by self-interested senders, Arieli et al. (2022) focuses on characterizing the sender's optimal value and extending concavification techniques.[14] Mahzoon (2022) focuses on the relationship between the receiver's welfare on the biases of uninformed intermediaries. Wu (2023) uses recursive concavification to study equilibrium information revelation. Lastly, Li & Norman (2021) studies how the number of senders affects equilibrium information revelation.

More generally, our model can be considered as one of Bayesian persuasion with multiple senders and receivers. In our model, however, the intermediary is not purely a sender or a receiver but both: she receives a persuasive message from a sender and uses it to design a new experiment to communicate with the receiver. As such, our framework presents substantial differences with the multi-receiver and multi-sender strategic communication literature (Kamenica, 2019; Milgrom & Roberts, 1986; Battaglini, 2002; Gentzkow & Kamenica, 2016a; Shin, 1998).[15] A critical difference between our work and the multi-receiver literature is that in our model the intermediary does not choose a payoff-relevant action but an experiment. Our framework is different to the multi-sender literature in two crucial ways. First, we consider a privately informed intermediary. Second, the intermediary's preferences coincide with the receiver's.

Finally, our model can be recast into one with receiver commitment without transfers, where the most notable application is mechanism design for optimal

[14] Although our method of proof is different, concavification techniques are key in Kamenica & Gentzkow's (2011) and much of the literature following it; they were first found in Aumann and Maschler's analysis of repeated games with incomplete information (Aumann et al., 1995).

[15] The literature about Bayesian persuasion with multiple receivers typically evolves around specific applications. Kamenica (2019) provides a good survey.

delegation, in the vein of Holmström (1978, 1984); Dessein (2002); Melumad & Shibano (1991); Alonso & Matouschek (2008). This is possible after noting that the intermediary and the receiver share the same preferences. As such, our model is strategically equivalent to one without an intermediary, where the receiver commits to a stochastic mapping from realizations of the sender's message and the private information into actions before such a message and private information are realized. Similarly, the optimal delegation literature studies the case where a sender has private information that can be communicated through messages, while the receiver commits to a deterministic or stochastic mapping from messages into actions. In contrast, we analyze how private information affects the efficacy of persuasive communication.

7. CONCLUDING REMARKS

In this paper, we have constructed a stylized model of persuasive communication with one sender (he), one receiver (she), and one privately-informed intermediary (she) representing the receiver's interests, and private information about the state of the world. A salient feature of our model is that the private information is unverifiable by the receiver.

With our model, we answered several relevant questions. Can private information or an intermediary improve the receiver's payoff over the payoff of her default action? Does private information's effectiveness depend on who communicates it to the receiver? Theorem 1 implies that private information alone cannot mitigate persuasive communication; an intermediary is needed. Moreover, Theorem 2 complements the finding by observing that private information must be accurate enough. Can private information or an intermediary change the sender's communication strategy? Proposition 2 through Proposition 5 imply that, except for very specific vacuous conditions, the sender modifies his behavior to account for the intermediary. However, the degree to which it benefits the receiver depends on the accuracy of the private information, as summarized by the AVIPI measure.

To derive our insights in a tractable way, we considered a simple binary-action and two-state model where the intermediary shares the receiver's preferences. With our simplifying assumptions, we can interpret our model as one that reflects situations where agents face simple accept-reject choices in a world where the state variable is either "good enough to grant acceptance" or not. Thus, we have focused on simple "direct" communication mechanisms instead of considering general mechanisms.

However, several aspects of the model can be generalized in relevant ways. Possible avenues involve determining whether focusing on direct experiments is without loss of generality, whether the results extend to arbitrary states and actions, and how the interaction between sender and intermediary changes when the latter's preferences are not perfectly aligned with the receiver's. Although these questions are of great importance, they are beyond the scope of this paper and some are addressed elsewhere (Murra-Anton et al., 2025).

A. PROOFS

Proof of Lemma 1. Let $\Gamma(\sigma_1, \sigma_2) = \sigma_1 s(a_1 \mid t) + \sigma_2 s(a_2 \mid t)$. $\Gamma(0,0) = 0$, $\Gamma(1,1) = 1$, and Γ is continuous. The intermediate value theorem implies that for any $y \in [0,1]$ there exist σ_1^y and σ_2^y such that $\Gamma(\sigma_1^y, \sigma_2^y) = y$. □

Proof of Theorem 1. The expected utility of the sender and the obedience constraint depend only on $\sigma(a \mid t)$. However, we can write $\sigma_t = \sigma(a \mid t) = \sum_{\hat{\sigma}} \sigma(a \mid \hat{s}, t) s(\hat{s} \mid t)$. Therefore, we can write the sender's optimization problem as

$$\max_{\sigma} \frac{1}{2}(\sigma_{t_g} + \sigma_{t_b})$$

$$\text{s.t. } \textit{Equation OC},$$

$$\text{s.t. } \sigma_t = \sum_{\hat{\sigma}} \sigma(a \mid \hat{s}, t) s(\hat{s} \mid t) \text{ for all } t.$$

Given that for any σ_t there exist $\sigma(a \mid \hat{s} = a_1, t)$ and $\sigma(a \mid \hat{s} = a_1, t)$ such that Equation IS is met, Equation IS does not represent an additional restriction to Equation OC in the sender's problem. Therefore, the solution to the problem is $\sigma_t = \sigma_t^{BP}$. Therefore, any $\sigma(a \mid \hat{s}, t)$ meeting Equation IS is sender-optimal. The player's expected utilities follow from the direct computation. □

Proof of Lemma 2. Let $V(\mu, \sigma)$ be the receiver's payoff from the experiment that follows from any given policies (σ, μ). The obedience constraint in Equation OC can be written as $V(\mu, \sigma) \geq 0$. Notice that $\mu(a_1 \mid \hat{s}, \hat{\sigma}) = 0$ for all $(\hat{s}, \hat{\sigma})$ is a feasible experiment, which we call μ_0. Under μ_0, $\mu_t = 0$ for all $t \in \{t_g, t_b\}$, and thus for any σ, $V(\mu_0, \sigma) = 0$. Then, for a given σ,

$\max_\mu V(\mu,\sigma) \geq V(\mu_0,\sigma) = 0$ and the intermediary's best response to any σ leads to an obedient experiment.[16] $\square$

Proof of Proposition 1. Fix $(\epsilon_g,\epsilon_b) \in [0,1]^2$ and $x \in (0,1)$. Consider the set $\Omega = [0,1]^6 \subset \mathbb{R}^6$ and the correspondence $\phi : \Omega \rightrightarrows \Omega$ defined through Equation 12 and Equation 14 as[17]

$$\phi = \begin{pmatrix} \mu^*(a_1 \mid t_g, a_1) \\ \mu^*(a_1 \mid t_g, a_2) \\ \mu^*(a_1 \mid t_b, a_1) \\ \mu^*(a_1 \mid t_b, a_2) \\ \sigma^*(a_1 \mid t_g) \\ \sigma^*(a_1 \mid t_b) \end{pmatrix}.$$

Given that $\mu^*(a_2 \mid \hat{s},\hat{\sigma}) = 1-\mu^*(a_1 \mid \hat{s},\hat{\sigma})$ and $\sigma^*(a_2 \mid t) = 1-\sigma^*(a_1 \mid t)$, a fixed point $\phi(\omega) = \omega \in \Omega$ is an equilibrium of the game. At least one such fixed point exists by Kakutani's Fixed Point Theorem because

1. Ω is nonempty, convex, and compact.
2. $\phi(\omega)$ is nonempty for every $\omega \in \Omega$.
3. $\phi(\omega)$ is convex-valued for every $\omega \in \Omega$ because each dimension is either single-valued or the convex set $[0,1]$.
4. ϕ is upper hemicontinuous because Ω is compact and ϕ is closed-graph. The latter follows from the fact that each dimension of ϕ is closed-graph itself. For example, consider $\mu^*(a_1 \mid t_g, a_1)$, since the rest of dimensions are analogous. For $(\sigma(a_1 \mid t_g), \sigma(a_1 \mid t_b))$ such that $\overline{v}(t_g, a_1) \neq 0$, $\mu^*(a_1 \mid t_g, a_1)$ is a continuous function, and hence it is upper hemicontinuous. For $(\sigma(a_1 \mid t_g), \sigma(a_1 \mid t_b)$ such that $\overline{v}(t_g, a_1) = 0$, $\mu^*(a_1 \mid t_g, a_1) \in \{0,1\}$ is part of the best-reply correspondence, hence, the image of any sequence in the domain converging to $(\sigma(a_1 \mid t_g), \sigma(a_1 \mid t_b))$ is part of and converges inside the image.

[16] Notice that the maximum is attained because $0 \leq \mu_t \leq 1$ for all t and $V(\mu,\sigma)$ is continuous in μ for all σ.

[17] Note well the slight abuse of notation. In general, μ^* is used as a specific value of the intermediary's policy, whereas in this proof it may be a set-valued function that allows for [0,1] as image.

□

Proof of Lemma 3. Equation 3 and Equation 7 imply that

$$V(\mu,\sigma)=\frac{1}{2}(x\mu_g-\mu_b)=$$

$$\frac{1}{2}\Bigg(\sum_{\hat{s},\hat{\sigma}}\mu(a_1\mid\hat{s},\hat{\sigma})s(\hat{s}\mid t_g)\sigma(\hat{\sigma}\mid t_g)x-\sum_{\hat{s},\hat{\sigma}}\mu(a_1\mid\hat{s},\hat{\sigma})s(\hat{s}\mid t_b)\sigma(\hat{\sigma}\mid t_b)\Bigg)=$$

$$\sum_{\hat{s},\hat{\sigma}}\mu(a_1\mid\hat{s},\hat{\sigma})\Bigg(\frac{1}{2}\big(x\sigma(\hat{\sigma}\mid t_g)s(\hat{s}\mid t_g)-\sigma(\hat{\sigma}\mid t_b)s(\hat{s}\mid t_b)\big)\Bigg)=$$

$$\sum_{\hat{s},\hat{\sigma}}\mu(a_1\mid\hat{s},\hat{\sigma})\overline{v}(\hat{s},\hat{\sigma}).$$

When $\overline{v}(\hat{s},\hat{\sigma})>0$, $\mu(a_1\mid\hat{s},\hat{\sigma})=1$ is maximal. Likewise, when $\overline{v}(\hat{s},\hat{\sigma})<0$, $\mu(a_1\mid\hat{s},\hat{\sigma})=0$ is maximal. Finally, $\overline{v}(\hat{s},\hat{\sigma})=0$, all $\mu(a_1\mid\hat{s},\hat{\sigma})\in[0,1]$ lead to the same payoff. As a result, the intermediary's best response to σ is

$$\mu^*(a_1\mid\hat{s},\hat{\sigma})=\begin{cases}1 & \text{if } \overline{v}(\hat{s},\hat{\sigma})>0,\\ \in[0,1] & \text{if } \overline{v}(\hat{s},\hat{\sigma})=0,\\ 0 & \text{if } \overline{v}(\hat{s},\hat{\sigma})<0.\end{cases}$$

Similarly, Equation 2 and Equation 7 imply that

$$U(\mu,\sigma)=\frac{1}{2}(\mu_g+\mu_b)=\frac{1}{2}\sum_{\hat{\sigma},\hat{s},t}\mu(a_1\mid\hat{s},\hat{\sigma})\sigma(\hat{\sigma}\mid t)s(\hat{s}\mid t)=$$

$$\frac{1}{2}\sum_{\hat{s},t}\big(\mu(a_1\mid\hat{s},a_1)\sigma(a_1\mid t)s(\hat{s}\mid t)+\mu(a_1\mid\hat{s},a_2)(1-\sigma(a_1\mid t))s(\hat{s}\mid t)\big)=$$

$$\frac{1}{2}\sum_{\hat{s},t}\big(\sigma(a_1\mid t)s(\hat{s}\mid t)(\mu(a_1\mid\hat{s},a_1)-\mu(a_1\mid\hat{s},a_2))+\mu(a_1\mid\hat{s},a_2)s(\hat{s}\mid t)\big)=$$

$$\frac{1}{2}\Bigg(\sum_{t}\sigma(a_1\mid t)\overline{\triangle}(t)+\sum_{\hat{s},t}\mu(a_1\mid\hat{s},a_2)s(\hat{s}\mid t)\Bigg).$$

When $\overline{\triangle}(t) > 0$, $\sigma(a_1 \mid t) = 1$ is maximal. Likewise, when $\overline{\triangle}(t) < 0$, $\sigma(a_1 \mid t) = 0$ is maximal. Finally, when $\overline{\triangle}(t) = 0$, all $\sigma(a_1 \mid t) \in [0,1]$ lead to the same payoff. As a result, the sender's best response to μ is

$$\sigma^*(a_1 \mid t) = \begin{cases} 1 & \text{if } \overline{\triangle}(t) > 0, \\ \in [0,1] & \text{if } \overline{\triangle}(t) = 0, \\ 0 & \text{if } \overline{\triangle}(t) < 0. \end{cases}$$

□

Proof of Theorem 2. To prove the first statement, we establish the following two steps:

Step 1: In any equilibrium (μ^*, σ^*), $V(\mu^*, \sigma^*) \geq \alpha(\epsilon_g, \epsilon_b)$.

Proof of Step 1: Fix any equilibrium experiment σ^* for the sender. In equilibrium, the experiment of the intermediary is such that $\mu^* \in argmax_{\mu} V(\mu, \sigma^*)$. Therefore, $V(\mu^*, \sigma^*) \geq V(\mu, \sigma^*)$ for all μ. Note that the strategy in which the intermediary ignores the sender, namely $\mu(a_1 \mid \hat{s}, a_1) = \mu(a_1 \mid \hat{s}, a_2)$ for all $\hat{s}$, is always feasible. Denote such a strategy by μ^A. We then have $V(\mu^A, \sigma^*) = \alpha(\epsilon_g, \epsilon_b) \leq V(\mu^*, \sigma^*)$.□

Step 2: In any equilibrium (μ^*, σ^*), $V(\mu^*, \sigma^*) > \alpha(\epsilon_g, \epsilon_b)$ is impossible.

Proof of Step 2: We show by contradiction that $V(\mu^*, \sigma^*) > \alpha(\epsilon_g, \epsilon_b)$ is impossible. To that end, we prove the following lemma:

Lemma 4. *Let (μ, σ) be any experiment profile with $V(\mu, \sigma) > \alpha(\epsilon_g, \epsilon_b)$. Then, there exists a neighborhood with radius $\gamma^* > 0$ around σ, denoted $N_{\gamma^*}(\sigma)$, such that for all $\sigma' \in N_{\gamma^*}(\sigma)$, $V(\mu, \sigma') > \alpha(\epsilon_g, \epsilon_b)$.*[18] *Thus, if (μ^*, σ^*) is an equilibrium such that $V(\mu^*, \sigma^*) > \alpha(\epsilon_g, \epsilon_b)$, the following are true:*

(i) $\overline{\triangle}(t) > 0$ *implies that* $\sigma^*_t = 1$.

(ii) $\overline{\triangle}(t) < 0$ *implies that* $\sigma^*_t = 0$.

Proof of Lemma 4. Consider an experiment profile (μ, σ). Equation 9 implies that $\overline{v}(\hat{s}, \hat{\sigma})$ is continuous in σ for all $(\hat{s}, \hat{\sigma})$, given s and x. Thus, Equation 11 implies that, given any fixed μ, $V(\mu, \sigma)$ is continuous in σ. Fixing μ, then,

[18] Formally, constrained to the probability simplex, $N_\gamma(\sigma) = \{\sigma' \in [0,1]^2 : ||\sigma - \sigma'|| < \gamma\}$, where $|| \cdot ||$ is the Euclidean distance.

for any $\theta > 0$, there is $\gamma_\theta > 0$ such that $|V(\mu, \sigma') - V(\mu, \sigma)| < \theta$ whenever $\sigma' \in N_{\gamma_\theta}(\sigma)$. This implies, in particular, that $V(\mu, \sigma) - \theta < V(\mu, \sigma')$. As $V(\mu, \sigma) > \alpha(\epsilon_g, \epsilon_b)$ by assumption, it is possible to set $\theta^* < V(\mu, \sigma)$ so that $V(\mu, \sigma') > \alpha(\epsilon_g, \epsilon_b)$. Thus, there exists $\gamma^* = \gamma_{\theta^*}$, such that for all $\sigma' \in N_{\gamma^*}(\sigma)$, $V(\mu, \sigma') > \alpha(\epsilon_g, \epsilon_b)$.

Let (μ^*, σ^*) be an equilibrium such that $V(\mu^*, \sigma^*) > \alpha(\epsilon_g, \epsilon_b)$. Suppose that $\overline{\triangle}(t) > 0$ and $\sigma_t^* < 1$. Denote by $-t$ the state that is not t. Step 1 and the first part of this lemma imply that it is possible to find $\sigma_t' > \sigma_t^*$ such that $\sigma' = (\sigma_t', \sigma_{-t}^*)$ is obedient given μ^*. However, since $\overline{\triangle}(t) > 0$, Equation 13 implies that $U(\mu^*, \sigma') > U(\mu^*, \sigma^*)$, a contradiction to (μ^*, σ^*) being an equilibrium. The case of $\overline{\triangle}(t) < 0$ is completely symmetric. □

In the remainder of the proof, we assume that (μ^*, σ^*) is an equilibrium such that $V(\mu^*, \sigma^*) > \alpha(\epsilon_g, \epsilon_b)$. We show that any possibility leads to a contradiction.

- Case 1: $\overline{\triangle}(t_g)\overline{\triangle}(t_b) > 0$. Suppose that $\overline{\triangle}(t_g) > 0$ and $\overline{\triangle}(t_b) > 0$. Then, Lemma 4 implies that $\sigma_g^* = \sigma_b^* = 1$. As a result, we can write

$$\overline{v}(\hat{s}, \hat{\sigma}) = \begin{cases} \underline{V}(\epsilon_g, \epsilon_b), & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_1), \\ \underline{V}'(\epsilon_g, \epsilon_b), & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_1), \\ 0, & \text{otherwise.} \end{cases}$$

Where $\underline{V}'(\epsilon_g, \epsilon_b) = \frac{1}{2}((1-\epsilon_g)x - \epsilon_b) \leq 0$. Therefore, the receiver's expected utility is

$$V(\mu^*, \sigma^*) = \mu^*(a_1 \mid t_g, a_1)\underline{V}(\epsilon_g, \epsilon_b) + \mu^*(a_1 \mid t_b, a_1)\underline{V}'(\epsilon_g, \epsilon_b).$$

As $\underline{V}'(\epsilon_g, \epsilon_b) \leq 0$, there are only two possibilities:[19] $\underline{V} > 0$ and $\underline{V}' < 0$, or $\underline{V} \leq 0$ and $\underline{V}' \leq 0$. In the first case, Equation 12 implies that $\mu^*(a_1 \mid t_g, a_1) = 1$ and $\mu^*(a_1 \mid t_b, a_1) = 0$, leading to $V(\mu^*, \sigma^*) = \underline{V}(\epsilon_g, \epsilon_b) = \alpha(\epsilon_g, \epsilon_b)$. A similar reasoning leads to $V(\mu^*, \sigma^*) = \alpha(\epsilon_g, \epsilon_b)$ when $\underline{V} \leq 0$ and $\underline{V}' \leq 0$. Therefore, $\overline{\triangle}(t_g) > 0$ and $\overline{\triangle}(t_b) > 0$ contradicts $V(\mu^*, \sigma^*) > \alpha(\epsilon_g, \epsilon_b)$.

The case of $\overline{\triangle}(t_g) < 0$ and $\overline{\triangle}(t_b) < 0$ is completely analogous. We conclude that $\overline{\triangle}(t_g)\overline{\triangle}(t_b) > 0$ is not possible in equilibrium.

[19] See Figure 1.

- Case 2: $\overline{\triangle}(t_g) = \overline{\triangle}(t_b) = 0$. Then, we can write

$$\overline{\triangle}(t_g) = \epsilon_g \triangle\mu(\hat{s} = t_g) + (1 - \epsilon_g)\triangle\mu(\hat{s} = t_b) = 0,$$

and

$$\overline{\triangle}(t_b) = (1 - \epsilon_b)\triangle\mu(\hat{s} = t_g) + \epsilon_b \triangle\mu(\hat{s} = t_b) = 0.$$

Notice that ϵ_g, $1-\epsilon_g$, ϵ_b, and $1-\epsilon_b$ cannot be zero simultaneously. Suppose $\epsilon_g \neq 0$, as the remaining cases are symmetric. Then, we can write $\triangle\mu(\hat{s} = t_g) = -\frac{1-\epsilon_g}{\epsilon_g}\triangle\mu(\hat{s} = t_b)$, and, thus, $0 = \frac{1-\epsilon_g-\epsilon_b}{\epsilon_g}\triangle\mu(\hat{s} = t_b)$.

Two sub-cases arise. In the first sub-case, $1 - \epsilon_g - \epsilon_b = 0$, which implies that $\underline{V}(\epsilon_g, \epsilon_b) \leq 0$, so $\alpha(\epsilon_g, \epsilon_b) = 0$. Moreover, we can write

$$\overline{v}(\hat{s}, \hat{\sigma}) = \frac{1}{2} \cdot \begin{cases} \epsilon_g(\sigma_g^* x - \sigma_b^*), & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_1), \\ (1 - \epsilon_g)(\sigma_g^* x - \sigma_b^*), & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_1), \\ \epsilon_g(x(1 - \sigma_g^*) - (1 - \sigma_b^*)), & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_2), \\ (1 - \epsilon_g)(x(1 - \sigma_g^*) - (1 - \sigma_b^*)), & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_2). \end{cases}$$

Note that $\sigma_g^* x - \sigma_b^* \geq 0$ implies that $x(1-\sigma_g^*) - (1-\sigma_b^*) < 0$ and $x(1-\sigma_g^*) - (1 - \sigma_b^*) \geq 0$ implies that $\sigma_g^* x - \sigma_b^* < 0$. Suppose $\epsilon_g \in (0, 1)$, as the case where $\epsilon_g \in \{0, 1\}$ is treated analogously. This creates four possibilities:

  - $\overline{v}(t_g, a_1) > 0, \overline{v}(t_b, a_1) > 0, \overline{v}(t_g, a_2) < 0$, and $\overline{v}(t_b, a_2) < 0$, which implies via Equation 12, $\mu^*(a_1 \mid t_g, a_1) = \mu^*(a_1 \mid t_b, a_1) = 1$ and $\mu^*(a_1 \mid t_g, a_2) = \mu^*(a_1 \mid t_b, a_2) = 0$. Therefore, $\triangle\mu(t_g) = 1 = \triangle\mu(t_b)$, and so $\overline{\triangle}(t_g) = 1$, a contradiction to the case assumption $\overline{\triangle}(t_g) = 0$.
  - $\overline{v}(t_g, a_1) < 0, \overline{v}(t_b, a_1) < 0, \overline{v}(t_g, a_2) > 0$, and $\overline{v}(t_b, a_2) > 0$, which leads to a similar contradiction.
  - $\overline{v}(t_g, a_1) = 0, \overline{v}(t_b, a_1) = 0, \overline{v}(t_g, a_2) < 0$, and $\overline{v}(t_b, a_2) < 0$, which implies via Equation 12, $\mu^*(a_1 \mid \hat{s}, a_1) \in [0, 1]$ for all $\hat{s}$ and $\mu^*(a_1 \mid \hat{s}, a_2) = 0$ for all $\hat{s}$. In any case, this implies via Equation 11, that $V(\mu^*, \sigma^*) = 0 = \alpha(\epsilon_g, \epsilon_b)$, a contradiction to the assumption $V(\mu^*, \sigma^*) > \alpha(\epsilon_g, \epsilon_b)$.
  - $\overline{v}(t_g, a_1) < 0, \overline{v}(t_b, a_1) < 0, \overline{v}(t_g, a_2) = 0$, and $\overline{v}(t_b, a_2) = 0$, which similarly leads to contradiction $V(\mu^*, \sigma^*) = \alpha(\epsilon_g, \epsilon_b)$.

All four cases lead to a contradiction in the sub-case $\epsilon_g + \epsilon_b - 1 = 0$. In the second sub-case, $\epsilon_g + \epsilon_b > 1$, so we must have $\epsilon_g > 0$. As $\frac{1-\epsilon_g-\epsilon_b}{\epsilon_g} \triangle\mu(\hat{s} = t_b) = 0$, then $\triangle\mu(\hat{s} = t_b) = 0$. Moreover, we had we $\triangle\mu(\hat{s} = t_g) = -\frac{1-\epsilon_g}{\epsilon_g} \triangle\mu(\hat{s} = t_b)$, which leads to $\triangle\mu(\hat{s} = t_g) = 0$. On the one hand, $\triangle\mu(\hat{s} = t_g) = 0$ implies that $\mu^*(a_1 \mid t_g, a_1) = \mu^*(a_1 \mid t_g, a_2)$, whereas $\triangle\mu(\hat{s} = t_b) = 0$ implies that $\mu^*(a_1 \mid t_b, a_1) = \mu^*(a_1 \mid t_b, a_2)$. As a result, using Equation 9 and Equation 11, we have that

$$V(\mu^*, \sigma^*) = \mu^*(a_1 \mid t_g, a_1)\underline{V}(\epsilon_g, \epsilon_b) - \mu^*(a_1 \mid t_b, a_1)\underline{V}'(\epsilon_g, \epsilon_b),$$

where $\underline{V}'(\epsilon_g, \epsilon_b) = \frac{1}{2}((1-\epsilon_g)x - \epsilon_b) \leq 0$.

As in Case 1, there are only two possibilities:[20] $\underline{V} > 0$ and $\underline{V}' < 0$, or $\underline{V} \leq 0$ and $\underline{V}' \leq 0$, both of which lead to $V(\mu^*, \sigma^*) = \alpha(\epsilon_g, \epsilon_b)$, a contradiction to the assumption of $V(\mu^*, \sigma^*) > \alpha(\epsilon_g, \epsilon_b)$.

We have shown that every possible case leads to a contradiction, so $\overline{\triangle}(t_g) = \overline{\triangle}(t_b) = 0$ is not possible in equilibrium.

- Case 3: $\overline{\triangle}(t_b) = 0$ and $\overline{\triangle}(t_g) \neq 0$ (as the opposite case is symmetric). We can write

$$\overline{\triangle}(t_b) = (1-\epsilon_b)\triangle\mu(\hat{s} = t_g) + \epsilon_b \triangle\mu(\hat{s} = t_b) = 0.$$

  If $\epsilon_b \in (0,1)$, we must have $\triangle\mu(\hat{s} = t_g) = \triangle\mu(\hat{s} = t_b) = 0$, so, as in the second sub-case of case 2, $\mu^*(a_1 \mid t_g, a_1) = \mu^*(a_1 \mid t_g, a_2)$ and $\mu^*(a_1 \mid t_b, a_1) = \mu^*(a_1 \mid t_b, a_2)$, which implies that $V(\mu^*, \sigma^*) = \alpha(\epsilon_g, \epsilon_b)$, a contradiction to the assumption of $V(\mu^*, \sigma^*) > \alpha(\epsilon_g, \epsilon_b)$. The remaining cases, $\epsilon_b \in \{0, 1\}$, analogously lead to the same contradiction. We have shown that $\overline{\triangle}(t_b) = 0$ and $\overline{\triangle}(t_g) \neq 0$ is not possible in equilibrium (and the opposite case is analogous).

- Case 4: $\overline{\triangle}(t_g)\overline{\triangle}(t_b) < 0$. Assume first that $\overline{\triangle}(t_g) > 0$ and $\overline{\triangle}(t_b) < 0$. Then, Lemma 4 implies that $\sigma_g^* = 1$ and $\sigma_b^* = 0$. Therefore, we can write

$$\overline{v}(\hat{s}, \hat{\sigma}) = \frac{1}{2} \cdot \begin{cases} \epsilon_g x, & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_1), \\ (1-\epsilon_g)x, & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_1), \\ -(1-\epsilon_b), & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_2), \\ -\epsilon_b, & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_2). \end{cases}$$

[20] See Figure 1.

Consider the case where $(\epsilon_g, \epsilon_b) \in (0,1)^2$. Therefore, Equation 12 implies that $\mu^*(a_1 \mid t_g, a_1) = \mu^*(a_1 \mid t_b, a_1) = 1$ and $\mu^*(a_1 \mid t_g, a_2) = \mu^*(a_1 \mid t_b, a_2) = 0$. Therefore, $\triangle\mu(\hat{s}) = 1$ for all $\hat{s}$, and so $\overline{\triangle}(t_b) > 0$, a contradiction. Any remaining combination of ϵ_g and ϵ_b leads to a similar contradiction. The case where $\overline{\triangle}(t_g) < 0$ and $\overline{\triangle}(t_b) > 0$ is analogous.

We have established that the assumption of $V(\mu^*, \sigma^*) > \alpha(\epsilon_g, \epsilon_b)$ always leads to a contradiction. Step 1 implies then that the only equilibrium possibility is $V(\mu^*, \sigma^*) = \alpha(\epsilon_g, \epsilon_b)$.

To conclude the proof, we next show the equivalence of (i) and (ii). Suppose that (ii) holds. Then for some $(\hat{s}, \hat{\sigma})$, $\overline{v}(\hat{s}, \hat{\sigma}) > 0$, and so Equation 12 implies that $\mu^*(a_1 \mid \hat{s}, \hat{\sigma}) = 1$. As $\mu^*(a \mid \hat{s}, \hat{\sigma})\overline{v}(\hat{s}, \hat{\mu}) \geq 0$ for all $(\hat{s}, \hat{\sigma})$, Equation 11 implies that $V(\mu^*, \sigma^*) > 0$, that is, (i) is valid.

We show that (i) implies (ii) by a contrapositive argument. Suppose that (ii) fails, so $\overline{v}(\hat{s}, \hat{\sigma}) \leq 0$ for all $(\hat{s}, \hat{\sigma})$. As $\mu^*(a_1 \mid \hat{s}, \hat{\sigma})\overline{v}(\hat{s}, \hat{\mu}) \geq 0$ and $\mu^*(a_1 \mid \hat{s}, \hat{\sigma}) \geq 0$ for all $(\hat{s}, \hat{\sigma})$, we conclude that $V(\mu^*, \sigma^*) = 0$, so (i) does not hold and the proof is complete. □

Proof of Corollary 1. Suppose $I(\epsilon_g, \epsilon_b) = \epsilon_g + \epsilon_b - 1 > 1 - x$. Then Theorem 2 implies that $V(\mu^*, \sigma^*) > 0$ because

$$\underline{V}(\epsilon_g, \epsilon_b) = \frac{1}{2}(x\epsilon_g - (1-\epsilon_b)) > \frac{1}{2}(1-\epsilon_g)(1-x) \geq 0.$$

To see that $I(\epsilon_g, \epsilon_b) > 1 - x$ is not necessary, suppose $x = 0.4$ and $\epsilon_g = \epsilon_b = 0.75$. Then $I(\epsilon_g, \epsilon_b) = 0.5 < 1 - x = 0.6$ but $\underline{V}(\epsilon_g, \epsilon_b) = \frac{1}{2}(x\epsilon_g - (1-\epsilon_b)) = 0.025 > 0$, hence $0 < \alpha(\epsilon_g, \epsilon_b) = V(\mu^*, \sigma^*)$ in any equilibrium. This completes part (i).

For part (ii), we proceed by a contra-positive argument. Suppose that $I(\epsilon_g, \epsilon_b) = 0$. Then, $\epsilon_g = (1 - \epsilon_b)$. Therefore,

$$\underline{V}(\epsilon_g, \epsilon_b) = \frac{1}{2}(x\epsilon_g - (1-\epsilon_b)) = \frac{1}{2}(x-1)\epsilon_g \leq 0,$$

, which implies that $\alpha(\epsilon_g, \epsilon_b)$. Then Theorem 2 implies that if (μ^*, σ^*) is an equilibrium, $V(\mu^*, \sigma^*) = 0$. We have shown that $I(\epsilon_g, \epsilon_b) = 0$ implies that $V(\mu^*, \sigma^*) = 0$, and hence $V(\mu^*, \sigma^*) > 0$ implies that $I(\epsilon_g, \epsilon_b) > 0$.

To show that $I(\epsilon_g, \epsilon_b) > 0$ is not sufficient, let $x = 0.1$ and $\epsilon_g = \epsilon_b = 0.51$. Then, $I(\epsilon_g, \epsilon_b) = 0.02 > 0$ but $\underline{V}(\epsilon_g, \epsilon_b) < 0$, so $0 = \alpha(\epsilon_g, \epsilon_b) = V(\mu^*, \sigma^*)$ in any equilibrium. □

Proof of Proposition 2. Part (i). To show the necessary condition, suppose that (μ^*, σ^*) is an equilibrium such that $\mu^*(a_1 \mid \hat{s}, a_1) = \mu^*(a_1 \mid \hat{s}, a_2)$ for all $\hat{s}$. If σ^* leads to $\overline{v}(\hat{s}, a_1)\overline{v}(\hat{s}, a_2) < 0$ for some $\hat{s}$, then Equation 12 implies that $\mu^*(a_1 \mid \hat{s}, a_1) = 0$ and $\mu^*(a_1 \mid \hat{s}, a_2) = 1$ or vice versa, a contradiction. Therefore, the only possibility is $\overline{v}(\hat{s}, a_1)\overline{v}(\hat{s}, a_2) \geq 0$ for all $\hat{s}$.

To show the sufficient condition, suppose that σ^* is an experiment such that $\overline{v}(\hat{s}, a_1)\overline{v}(\hat{s}, a_2) \geq 0$ for all $\hat{s}$. Only four possibilities arise, for a given $\hat{s}$:

- $\overline{v}(\hat{s}, a_1) = 0$. In this case, any $\mu(a_1 \mid \hat{s}, a_1) \in [0, 1]$ is optimal for the intermediary. Let $\mu^*(a_1 \mid \hat{s}, a_2)$ be any best response of the intermediary when $(\hat{s}, a_2)$ is observed. Then, we can set $\mu^*(a_1 \mid \hat{s}, a_1) = \mu^*(a_1 \mid \hat{s}, a_2)$, which leads to $\triangle\mu(\hat{s}) = 0$.

- $\overline{v}(\hat{s}, a_2) = 0$. In this case, any $\mu(a_1 \mid \hat{s}, a_2) \in [0, 1]$ is optimal for the intermediary. Let $\mu^*(a_1 \mid \hat{s}, a_1)$ be any best response of the intermediary when $(\hat{s}, a_1)$ is observed. Then, we can set $\mu^*(a_1 \mid \hat{s}, a_2) = \mu^*(a_1 \mid \hat{s}, a_1)$, which leads to $\triangle\mu(\hat{s}) = 0$.

- $\overline{v}(\hat{s}, a_1) < 0$ and $\overline{v}(\hat{s}, a_2) < 0$. In this case, the intermediary's only best response is $\mu(a_1 \mid \hat{s}, a_1) = \mu(a_1 \mid \hat{s}, a_2) = 0$, leading to $\triangle\mu(\hat{s}) = 0$.

- $\overline{v}(\hat{s}, a_1) > 0$ and $\overline{v}(\hat{s}, a_2) > 0$. In this case, the intermediary's only best response is $\mu(a_1 \mid \hat{s}, a_1) = \mu(a_1 \mid \hat{s}, a_2) = 1$, leading to $\triangle\mu(\hat{s}) = 0$.

We conclude that when σ^* is such that $\overline{v}(\hat{s}, a_1)\overline{v}(\hat{s}, a_2) \geq 0$, $\mu^*(a_1 \mid \hat{s}, a_1) = \mu^*(a_1 \mid \hat{s}, a_2)$ for all $\hat{s}$ is a best response for the intermediary. It is left to check that σ^* is a best response to μ^* above. In any possible case, as described above, the only possibility is $\triangle\mu(\hat{s}) = 0$ for all $\hat{s}$. Thus, $\overline{\triangle}(t) = 0$ for all t. Therefore, the sender is indifferent to all $\sigma_t \in [0, 1]$, for all t. Therefore, σ^* is a best response to μ^* above.

Part (ii). Trivially, if $\sigma_g^* = \sigma_b^* = 0$, the necessary and sufficient condition in part (i) is satisfied, independently of ϵ_g and ϵ_b. This is not the only type of equilibrium in which the intermediary always ignores the sender. $\sigma_g^* = 1$ and $\sigma_b^* = x$ when $\epsilon_g = 1 - \epsilon_b$ is a different possibility.

Part (iii). Suppose in equilibrium $\mu^*(a_1 \mid \hat{s}, a_1) = \mu^*(a_1 \mid \hat{s}, a_2) = \mu^*(\hat{s})$ for all $\hat{s}$. As a result, we can write from Equation 7,

$$\mu_g = \mu^*(t_g)\epsilon_g + \mu^*(t_b)(1 - \epsilon_g),$$

and

$$\mu_b = \mu^*(t_g)(1-\epsilon_b) + \mu^*(t_b)\epsilon_b.$$

Using Equation 3, we can write the expected utility of following the recommendation a_1 as

$$V(\mu^*,\sigma^*) = \frac{1}{2}(\mu_g x - \mu_b) = \frac{1}{2}(\mu^*(t_g)[x\epsilon_g - (1-\epsilon_b)] + \mu^*(t_b)[x(1-\epsilon_g) - \epsilon_b]) =$$

$$\mu^*(t_g)\underline{V}(\epsilon_g,\epsilon_b) + \mu^*(t_b)\underline{V}'(\epsilon_g,\epsilon_b) \geq 0,$$

where $\underline{V}'(\epsilon_g,\epsilon_b) = \frac{1}{2}((1-\epsilon_g)x - \epsilon_b) \leq 0$, and so the inequality follows from Theorem 2, which implies that $V(\mu^*,\sigma^*) = \alpha(\epsilon_g,\epsilon_b) \geq 0$. The above equation implies that the intermediary's best response is as follows: $\mu^*(t_g) = 1$ when $\underline{V} > 0$, $\mu^*(t_g) = 0$ when $\underline{V} < 0$, $\mu^*(t_g) \in [0,1]$ when $\underline{V} = 0$, and similarly for t_b and $\underline{V}'$. □

Proof of Proposition 3. For the necessary condition, suppose that (μ^*,σ^*) is an equilibrium where the intermediary always passes the sender's message along. Then, $\mu^*(a_1 \mid \hat{s}, a_1) = 1$ and $\mu^*(a_1 \mid \hat{s}, a_2) = 0$ for all $\hat{s}$. Consequently, $\triangle\mu^*(\hat{s}) = 1$ for all $\hat{s}$, and so $\overline{\triangle}(t) = 1$ for all t.

Equation 7 and μ^* imply that $\mu_g^* = \sigma_g^*$ and $\mu_b^* = \sigma_b^*$, so Equation 3 implies that $V(\mu^*,\sigma^*) = \frac{1}{2}(\sigma_g^* x - \sigma_b^*)$. Finally, Theorem 2 implies that the sender chooses his best response experiment σ^* to solve

$$\begin{aligned} \max_{\sigma_g,\sigma_b} \quad & \sigma_g + \sigma_b \\ \text{s.t.} \quad & V(\mu^*,\sigma) = \frac{1}{2}\left(\sigma_g x - \sigma_b\right) = \alpha(\epsilon_g,\epsilon_b). \end{aligned} \tag{15}$$

There are two relevant cases, $\alpha(\epsilon_g,\epsilon_b) > 0$ and $\alpha(\epsilon_g,\epsilon_b) = 0$.

- Case 1, $\alpha(\epsilon_g,\epsilon_b) > 0$. Consider the case where $\alpha(\epsilon_g,\epsilon_b) = \underline{V}(\epsilon_g,\epsilon_b) = \frac{1}{2}(\epsilon_g x - (1-\epsilon_b)) > 0$. The solution to the sender's problem is $\sigma_g^* = 1$ and $\sigma_b^* = x - \underline{V}(\epsilon_g,\epsilon_b) < x$. Then, since $\underline{V}(\epsilon_g,\epsilon_b) > 0$ implies that $\epsilon_g + \epsilon_b - 1 < 0$, we have $\overline{v}(t_b, a_1) = \frac{1}{2}x(1-\epsilon_g-\epsilon_b) < 0$. However, this implies through Equation 12 that $\mu^*(a_1 \mid t_b, a_1) = 0$, a contradiction.

- Case 2, $\alpha(\epsilon_g,\epsilon_b) = 0$. The sender's problem coincides with the problem in Equation BP, and the unique optimal experiment is $\sigma_g^* = 1$ and $\sigma_b^* = x$.

Therefore, we have

$$\overline{v}(\hat{s}, \hat{\sigma}) = \begin{cases} \frac{x}{2}(\epsilon_g + \epsilon_b - 1), & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_1), \\ \frac{x}{2}(1 - \epsilon_g - \epsilon_b), & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_1), \\ -\frac{(1-x)}{2}(1 - \epsilon_b), & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_2), \\ -\frac{(1-x)}{2}\epsilon_b, & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_2). \end{cases}$$

We have that $\overline{v}(\hat{s}, a_2)$ for all $\hat{s}$, which supports an intermediary's best response of $\mu^*(a_1 \mid \hat{s}, a_2) = 0$ according to Equation 12. If $I(\epsilon_g, \epsilon_b) > 0$, $\overline{v}(t_g, a_1)\overline{v}(t_b, a_1) < 0$, which leads to a best response of $\mu^*(a_1 \mid \hat{s}, a_1) = 0$ for some $\hat{s}$, a contradiction. Finally, if $I(\epsilon_g, \epsilon_b) = 0$, $\overline{v}(\hat{s}, a_1) = 0$ for all $\hat{s}$, which supports an intermediary's best response of $\mu^*(a_1 \mid \hat{s}, a_2) = 1$ according to Equation 12.

From the two cases above, we conclude that if (μ^*, σ^*) is an equilibrium such that $\mu^*(a_1 \mid \hat{s}, a_1) = 1$ and $\mu^*(a_1 \mid \hat{s}, a_2) = 0$ for all $\hat{s}$, we must have $I(\epsilon_g, \epsilon_b) = 0$, $\sigma_g^* = 1$ and $\sigma_b^* = x$.

To show the sufficient condition, suppose that $I(\epsilon_g, \epsilon_b) = 0$, $\sigma_g^* = 1$, and $\sigma_b^* = x$. Then, we have

$$\overline{v}(\hat{s}, \hat{\sigma}) = \begin{cases} 0, & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_1), \\ 0, & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_1), \\ -\frac{(1-x)}{2}(1 - \epsilon_b) \leq 0, & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_2), \\ -\frac{(1-x)}{2}\epsilon_b \leq 0, & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_2). \end{cases}$$

Hence, a best response to the sender's strategy according to Equation 12 is $\mu^*(a_1 \mid \hat{s}, a_1) = 1$ and $\mu^*(a_1 \mid \hat{s}, a_2) = 0$ for all $\hat{s}$. Under μ^*, the sender's problem is in Equation 15, which is solved by $\sigma_g^* = 1$ and $\sigma_b^* = x$. We conclude that if $I(\epsilon_g, \epsilon_b) = 0$, $\sigma_g^* = 1$, and σ_b^*, there exists an equilibrium where the intermediary always passes along the sender's recommendation. □

Proof of Proposition 4. In the main text. □

Proof of Proposition 5. Part (i) follows directly from Proposition 3.

For part (ii), assume that $\sigma_g^* = 1$ and $\sigma_b^* = x$ and let $I(\epsilon_g, \epsilon_b) = \epsilon_g + \epsilon_b - 1 > 0$. Then, we can compute the following:

$$\overline{v}(\hat{s}, \hat{\sigma}) = \begin{cases} \frac{x}{2}(\epsilon_g + \epsilon_b - 1), & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_1), \\ \frac{x}{2}(1 - \epsilon_g - \epsilon_b), & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_1), \\ -\frac{(1-x)}{2}(1 - \epsilon_b), & \text{if } (\hat{s}, \hat{\sigma}) = (t_g, a_2), \\ -\frac{(1-x)}{2}\epsilon_b, & \text{if } (\hat{s}, \hat{\sigma}) = (t_b, a_2). \end{cases}$$

As $\epsilon_g + \epsilon_b - 1 > 0$, we have that $\overline{v}(t_g, a_1) > 0$ and $\overline{v}(t_b, a_1) < 0$, so Equation 12 implies that $\mu^*(a_1 \mid t_g, a_1) = 1$ and $\mu^*(a_1 \mid t_b, a_1) = 0$. We have three sub-cases:

- If $\epsilon_b \in (0, 1)$, we have $\overline{v}(\hat{s}, a_2) < 0$ for all $\hat{s}$, so Equation 12 implies that $\mu^*(a_1 \mid \hat{s}, a_2) = 0$. This implies that $\triangle\mu^*(t_g) = 1$ and $\triangle\mu^*(t_b) = 0$. Therefore, $\overline{\triangle}(t_b) = (1 - \epsilon_b) > 0$. However, Equation 14 implies that $\sigma_b^* = 1$, a contradiction.

- If $\epsilon_b = 0$, a similar reasoning implies that $\overline{\triangle}(t_b) = (1 - \epsilon_b) > 0$, which leads to $\sigma_b^* = 1$, a contradiction.

- If $\epsilon_b = 1$, we must have $\epsilon_g > 0$ because $\epsilon_g + \epsilon_b - 1 > 0$. Moreover, we have $\overline{v}(t_g, a_2) = 0$ and $\overline{v}(t_b, a_2) < 0$, so Equation 12 implies that $\mu^*(a_1 \mid t_g, a_2) \in [0, 1]$ and $\mu^*(a_1 \mid t_b, a_2) = 0$. Then, we have $\triangle\mu^*(t_g) \geq 0$ and $\triangle\mu^*(t_b) = 0$. Hence, $\overline{\triangle}(t_g) \geq 0$ and $\overline{\triangle}(t_b) = 0$, which imply, via Equation 14, that $\sigma_g^* = 1$ and $\sigma_b^* = x$ is consistent with equilibrium behavior. The payoffs and the investment probability follow from the direct computation of Equation 11 and Equation 13.

□

B. APPLICATIONS

- A pharmaceutical company (sender) designs a clinical trial (communication) to persuade the FDA (receiver) to approve a new drug, which can be dangerous (bad) or safe (good) for the general population. The FDA typically designates a technical committee (intermediary) to review related scientific evidence (private information) to make a recommendation.

- A supervisor (sender) writes a letter of recommendation advocating for the promotion of an employee whose productivity can be superb (good) or average (bad). The Human Resources department (mediator) reviews the recommendation letter, along with independent performance metrics (private information). HR then recommends the promotion to the company board (receiver).

- A research analyst (sender) produces an investment report (communication) on a given asset that can be safe (good) or risky (bad). The firm's risk management team (mediator) integrates the report with other market data and risk assessments (private information). Finally, the risk management team makes a recommendation to the portfolio manager (receiver).

- A lobby group (sender) submits a policy proposal (communication) to influence regulatory changes in a social program that can be obsolete (bad) or functional (good). A regulatory agency (mediator) receives the proposal and collects related data (private information). The regulatory agency makes a recommendation to Congress (receiver).

- A seller (sender) submits product descriptions, tests, evidence, and certifications (communication) to an online platform like Amazon (intermediary) about a product that can be good or bad. The platform aggregates the seller's information with online reviews and other market metrics (private information). The platform makes a purchase recommendation to consumers (receiver).

References

Alonso, R., & Matouschek, N. (2008). Optimal delegation. *Review of Economic Studies*, *75*(1), 259–293.

Ambrus, A., Azevedo, E. M., & Kamada, Y. (2013). Hierarchical cheap talk. *Theoretical Economics*, *8*(1), 233–261.

Arieli, I., Babichenko, Y., & Sandomirskiy, F. (2022). Bayesian persuasion with mediators. *arXiv preprint arXiv:2203.04285*.

Aumann, R. J., Maschler, M., & Stearns, R. E. (1995). *Repeated Games with Incomplete Information*. MIT press.

Battaglini, M. (2002). Multiple referrals and multidimensional cheap talk. *Econometrica*, *70*(4), 1379–1401.

Bergemann, D., & Morris, S. (2016). Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics*, *11*(2), 487–522.

Bergemann, D., & Morris, S. (2019). Information design: A unified perspective. *Journal of Economic Literature*, *57*(1), 44–95.

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, 1431–1451.

Dessein, W. (2002). Authority and communication in organizations. *Review of Economic Studies*, *69*(4), 811–838.

Forges, F. (1986). An approach to communication equilibria. *Econometrica*, 1375–1385.

Gentzkow, M., & Kamenica, E. (2016a). Competition in persuasion. *Review of Economic Studies*, *84*(1), 300–322.

Gentzkow, M., & Kamenica, E. (2016b). A Rothschild-Stiglitz approach to Bayesian persuasion. *American Economic Review*, *106*(5), 597–601.

Holmström, B. (1978). *On incentives and control in organizations*. Stanford University.

Holmström, B. (1984). On the theory of delegation. In M. Boyer & R. Kihlstrom (Eds.), *Bayesian Models in Economic Theory* (pp. 115–41). New York, NY: North-Holland.

Ivanov, M. (2010). Communication via a strategic mediator. *Journal of Economic Theory*, *145*(2), 869–884.

Kamenica, E. (2019). Bayesian persuasion and information design. *Annual Review of Economics*, *11*, 249–272.

Kamenica, E., & Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, *101*(6), 2590–2615.

Kolotilin, A., Mylovanov, T., Zapechelnyuk, A., & Li, M. (2017). Persuasion of a privately informed receiver. *Econometrica*, *85*(6), 1949–1964.

Kosenko, A. (2020). Mediated persuasion. *arXiv preprint arXiv:2012.00098*.

Li, F., & Norman, P. (2021). Sequential persuasion. *Theoretical Economics*, *16*(2), 639–675.

Mahzoon, M. (2022). Hierarchical Bayesian persuasion: Importance of vice presidents. *arXiv preprint arXiv:2204.05304*.

Melumad, N. D., & Shibano, T. (1991). Communication in settings with no transfers. *RAND Journal of Economics*, 173–198.

Milgrom, P., & Roberts, J. (1986). Relying on the information of interested parties. *RAND Journal of Economics*, 18–32.

Murra-Anton, Z., Rubbini, G., & Serrano, R. (2025). The recommendation principle in information design. *Working Paper, available at SSRN 5137988*.

Perez-Richet, E., & Skreta, V. (2022). Test design under falsification. *Econometrica*, *90*(3), 1109–1142.

Shin, H. S. (1998). Adversarial and inquisitorial procedures in arbitration. *RAND Journal of Economics*, 378–405.

Wu, W. (2023). Sequential Bayesian persuasion. *Journal of Economic Theory*, *214*, 105763.

Journal of Mechanism and Institution Design
ISSN: 2399-844X (Print), 2399-8458 (Online)
DOI:10.22574/jmid.2025.12.002

EQUILIBRIUM CONDITIONS FOR CATCH-22 SITUATIONS

Joshua S. Gans, University of Toronto, Canada and NBER, USA
joshua.gans@utoronto.ca

ABSTRACT

This paper examines whether Catch-22 situations persist as equilibrium phenomena based on Gatekeepers' preferences rather than arbitrary rules. We model a game-theoretic scenario inspired by Heller's paradox, featuring a Requester (either Sane or Insane) who signals before a Gatekeeper decides to grant or deny their request. Our analysis identifies conditions where Catch-22 situations emerge as equilibria and evaluates their efficiency properties. Results demonstrate that such situations can arise naturally as efficient responses by Gatekeepers facing asymmetric information. The model is extended to labor markets, where "experience required for jobs, but jobs required for experience" creates similar dynamics.

Keywords: Catch-22 situation, asymmetric information, dilemmas.

JEL Classification Numbers: C72, D82, J24, D86.

> *"You mean there's a catch?"*
>
> *"Sure there's a catch," Doc Daneeka replied. "Catch-22. Anyone who wants to get out of combat duty isn't really crazy."*
>
> *There was only one catch and that was Catch-22, which specified that a concern for one's own safety in the face of dangers that were real and immediate was the process of a rational mind. Orr was crazy and could be grounded. All he had to do was ask; and as soon as he did, he would no longer be crazy and would have to fly more missions. Orr would be crazy to fly more missions and*

The author declares there are no conflicts of interest. I thank ChatGPT-o1-pro for helpful research assistance. Any errors are my own.

> *sane if he didn't, but if he was sane, he had to fly them. If he flew them, he was crazy and didn't have to; but if he didn't want to, he was sane and had to. Yossarian was moved very deeply by the absolute simplicity of this clause of Catch-22 and let out a respectful whistle.* (Heller, 1961)

1. INTRODUCTION

JOSEPH Heller's classic *Catch-22* portrays a paradox: a pilot can only be relieved from flying dangerous missions if they are insane. But any pilot who requests relief, by that very action, reveals a concern for self-preservation, implying sanity. Thus, to qualify for relief, one must never request it, but not requesting fails to demonstrate insanity.

This is a canonical Catch-22 situation: an agent cannot access a rule because complying with the rule prevents them from achieving their desired outcome. In popular culture, it presents a type of 'no win' situation or dilemma but of a particular kind. Specifically, the rule is set by those in power and ultimately leaves others with no real choice. It connotes a situation where such rules result in oppression, and their rationale subverts outcomes other than those pushed by those in power. Consequently, Catch-22 situations have been argued to be present in real-world situations, e.g., "to receive a loan, you need a credit score, but to obtain a credit score, you need a loan."

Economists, as they are wont, nay have a duty, to do, might wonder: where did these rules come from in the first place? Are they a mere expression of power dressed up in a rationale or did they express the preferences of those who imposed them? This applies in Heller's scenario. There, the authority (Gatekeeper) expresses their own preferences: they do not want to ground rational (Sane) pilots who should remain in action, but they do want insane pilots to be kept from flying. This structure suggests we must model not only the paradoxical condition but also the payoffs to both sides. Only then can we understand why the rule is in place, what its function is and whether it achieves that function. Only then do we have a theory to be put against the alternative notion that the rule exists because of guile, comedic intent or existential angst.

We formalize the game as follows. There are two players: a Requester, who can be one of two types (Sane or Insane) and a Gatekeeper. The Requester chooses to send a message (Request or Not Request) before the Gatekeeper decides whether to grant their request or not. The Requester's sanity is defined by their preferences: Sane types prefer A (not to fly) to B (to fly), while Insane types prefer B to A; indeed, their preferences define their insanity. The Gatekeeper's preferences run counter: the Gatekeeper wants Sane types to get B and Insane types to get A. In Heller's terms, the reason the Gatekeeper does not want insane agents to fly is that they would prefer them not to fly.

This game allows us to explore the two characteristics that underpin a Catch-22 situation. First, the agent subject to the rule *observe*s its futility in practice. In Heller's case, their request not to fly will always be denied. Second, the Gatekeeper draws clear *inferences* from the Requester of their type based on the message they send (i.e., the request they do or do not make). In analysing this game, we look for equilibrium outcomes that are observationally a Catch-22 situation in that the rule is applied and also ones that are underpinned by the beliefs associated with a Catch-22 situation.

Section 2 sets the game up, assuming that the Requester moves first and the Gatekeeper responds. Section 3 then solves for the equilibrium outcome. It is demonstrated that a Catch-22 situation – that is, where a Sane type who requests not to fly will have to fly – only arises if the portion of Sane types in the population is relatively high with a threshold related to the Gatekeeper's costs of allowing an insane person to fly or preventing a sane person from flying. When there are many Insane types, the Gatekeeper is better off allowing sane types to request and fulfil their wish not to fly and barring anyone else from flying on account of them being insane. The reverse timing, where the Gatekeeper commits to a rule ex ante, is shown to give rise to the same equilibrium outcomes.

In Section 4, we then turn to consider a more economically relevant Catch-22 (conjectured) situation: "You can't get a job without experience, but you can't get experience without a job." This is done in the context of a labour market where getting a job gives you experience and the ability to perform (and earn) based on higher productivity.

A final section concludes with thoughts about this general class of Catch-22 situations with a view to discouraging any others from wasting more time pursuing this line of research.

There is a small literature on Catch-22 as it relates to economics. It pops up to motivate a few policy papers, including Colgan & Quinlin (1997), de Rouffignac (2000), Flood & Toner (1997), Brams (1999) and Stein (1999).

Brams & Jones (1999) explores some game theoretic explorations in the context of 2x2 games and power while Carmichael (2002) considers a potential Catch-22 situation in a signalling game where sending the desired message to one receiver results in the wrong message to the other and vice versa. None of these papers, however, have tackled the Catch-22 situation as spelled out by Heller (1961) head-on nor explored its plausibility for that case.[1]

2. MODEL SETUP

There are two players: a Requester and a Gatekeeper. The Requester is either Sane (S) or Insane (I). Let $p \in (0,1)$ be the probability that the Requester is Insane, and $1-p$ the probability that the Requester is Sane. There is asymmetric information as the Requester knows their type while the Gatekeeper does not. The Gatekeeper only knows p.

The Requester chooses between two messages: R (Request) or $\neg R$ (Not Request). After observing the Requester's action, the Gatekeeper chooses between G (Grant) or D (Deny). If the Gatekeeper chooses G, the outcome is A; if D, the outcome is B.[2] Thus, the timing of the game is as follows:

1. Nature chooses the Requester's type: S with probability $1-p$ or I with probability p.
2. The Requester chooses R or $\neg R$.
3. The Gatekeeper observes the Requester's action and chooses G or D.

The preferences of each player are as follows:

- **Requester's Preference**s: Let $r_i(a)$ be the payoff a Requester of type $i \in \{S,I\}$ receives with an outcome $a \in \{A,B\}$. It is assumed that $r_S(A) > r_S(B)$ (a Sane Requester prefers A over B) and that $r_I(A) < r_I(B)$ (an Insane Requester prefers B over A.

[1] There is some work in logical philosophy on Catch-22 situations that argues that the proposed rule and inference are vacuous and uninformative (Goldstein, 2004). Here, I consider a game that allows for non-arbitrary rules, and any rule formation is consistent with incentives and beliefs. Thus, it provides a different perspective than this literature.

[2] An alternative timing where the Gatekeeper first commits to a rule is considered below with no change in the equilibrium outcomes.

- **Gatekeeper's Preferences**: Let $g_I(a)$ be the payoff to the Gatekeeper if the Requester is of type $i \in \{S, I\}$ and receives an outcome $a \in \{A, B\}$. For a Sane Requester, $g_S(A) < g_S(B)$ if the Gatekeeper prefers B over A. Otherwise, $g_S(A) > g_S(B)$ if the Gatekeeper prefers A over B.

Thus, sanity is defined by which outcome the Requester prefers: a Sane Requester wants A, an Insane one wants B.

3. EQUILIBRIUM OUTCOMES

We examine perfect Bayesian equilibria of the game; that is, a profile of strategies and beliefs where each player's strategy is a best response given their beliefs at each stage of the game, and beliefs are consistent with Bayes' rule whenever possible. For each type of Requester, a strategy is a probability ($\beta_S \equiv P(R|S)$ for Sane, $\beta_I \equiv P(R|I)$ for Insane) of choosing R and, for the Gatekeeper, a strategy is probabilities, $(\alpha_R, \alpha_{\neg R})$ of choosing G after R or $\neg R$. In equilibrium, (i) given the Gatekeeper's strategy, each type of Requester maximises their expected utility by choosing R or $\neg R$; (ii) given the Requesters' strategies, the Gatekeeper chooses G or D to maximise expected utility after observing R or $\neg R$ and (iii) the Gatekeeper's beliefs, $P(I|R)$ and $P(I|\neg R)$ are updated using Bayes' rule:

$$P(I|R) = \frac{p\beta_I}{p\beta_I + (1-p)\beta_S}, \quad P(I|\neg R) = \frac{p(1-\beta_I)}{p(1-\beta_I) + (1-p)(1-\beta_S)}$$

In some cases, if certain actions are played with probability zero, off-path beliefs must be specified.

Working backwards, after seeing R, the Gatekeeper's expected payoffs from choosing G and D are respectively:

$$E[G|R] = P(I|R)g_I(A) + (1 - P(I|R))g_S(A).$$

$$E[D|R] = P(I|R)g_I(B) + (1 - P(I|R))g_S(B).$$

The Gatekeeper grants after R if $E[G|R] > E[D|R]$. Similarly for $\neg R$.

Define $\Delta_I = g_I(A) - g_I(B) > 0$ and $\Delta_S = g_S(A) - g_S(B) < 0$. Then:

$$E[G|R] - E[D|R] = P(I|R)\Delta_I + (1 - P(I|R))\Delta_S.$$

It is straightforward to see that the Gatekeeper will choose G rather than D following a request (R) if:

$$P(I|R) > \frac{-\Delta_S}{\Delta_I - \Delta_S}.$$

and will choose G rather than D following no request ($\neg R$) if:

$$P(I|\neg R) > \frac{-\Delta_S}{\Delta_I - \Delta_S}.$$

Otherwise, the Gatekeeper will be indifferent.

Turning to the Requester's choice, given an expectation that the Gatekeeper chooses $(\alpha_R, \alpha_{\neg R})$, the Sane type chooses R if it yields a higher expected payoff, that is:

$$r_S(R) = \alpha_R r_S(A) + (1-\alpha_R) r_S(B), \quad r_S(\neg R) = \alpha_{\neg R} r_S(A) + (1-\alpha_{\neg R}) r_S(B).$$

Note that Sane is indifferent if $\alpha_R = \alpha_{\neg R}$. Similarly, for an Insane type:

$$r_I(R) = \alpha_R r_I(A) + (1-\alpha_R) r_I(B), \quad r_I(\neg R) = \alpha_{\neg R} r_I(A) + (1-\alpha_{\neg R}) r_I(B).$$

Indifference implies $\alpha_R = \alpha_{\neg R}$ as well. So if both types must be simultaneously indifferent, $\alpha_R = \alpha_{\neg R} = \alpha$. Otherwise, if $\alpha_R \neq \alpha_{\neg R}$, both Sane and Insane Requesters will find a pure strategy optimal. If $\alpha_R > (<) \alpha_{\neg R}$, Sane chooses R ($\neg R$) and Insane chooses $\neg R$ (R).

Given this, we can now prove the following:

Proposition 1. *Consider the game where the Requester moves first. Let*

$$q \equiv \frac{-\Delta_S}{\Delta_I - \Delta_S} = \frac{g_S(B) - g_S(A)}{(g_I(A) - g_I(B)) + (g_S(B) - g_S(A))}.$$

1. *If $p < q$, the unique equilibrium outcome has the Gatekeeper always deny.*

2. *If $p > q$, the unique equilibrium outcome has the Gatekeeper always granting.*

3. *If $p = q$, there is a continuum of equilibria where $\beta_S = \beta_I$ and $\alpha_R = \alpha_{\neg R}$.*

All proofs are in the appendix. When the probability of insanity p is low (less than q), the Gatekeeper expects mostly Sane types and, thus, prefers forcing B by denying. When p is high ($p > q$), expecting more Insane types, the Gatekeeper grants to achieve outcome A. Only at the exact balance point $p = q$ is the Gatekeeper genuinely uncertain and indifferent, leading to a mixed strategy outcome that is, arguably, non-generic.[3]

This demonstrates that a Catch-22 situation whereby the Gatekeeper always, observationally, denies a request for A from a Requester only arises if p is sufficiently low (less than q). This is because, in that case, the Gatekeeper's error from having an Insane person with outcome B (Δ_I) is lower relative to the error in having a Sane person with outcome A (Δ_S). Note, however, while the Gatekeeper is acting *as if* it is denying because it is inferring that only a Sane person would make a request R, the Gatekeeper's actual equilibrium beliefs that the person making the request is Sane is arbitrary. The Gatekeeper could hold those beliefs or not. It does not matter for the outcome. This is the classic babbling equilibrium outcome from Crawford & Sobel (1982).

More critically, the equilibrium outcome is driven by efficiency considerations. Absent an ability to infer an agent's type from cheap talk messages, the Gatekeeper's action reflects a balancing of the types of errors they could make. Thus, the rule is not arbitrary in this model but reflects the preferences of the Gatekeeper. Specifically, Proposition 1 shows that if there is a large volume of Insane types, the Gatekeeper would not employ a Catch-22-like rule.

Now consider what happens if the Gatekeeper commits to $(\alpha_R, \alpha_{\neg R})$ before the Requester chooses R or $\neg R$. The Requester, observing $(\alpha_R, \alpha_{\neg R})$, picks the message that maximises their payoff. If $p < q$, any attempt to equalise payoffs by mixing is exploited by Requesters. The Gatekeeper ends up best off by choosing a pure strategy. The same logic applies for $p > q$. Thus, we get the same outcomes: always deny if $p < q$, always grant if $p > q$. If $p = q$, the Gatekeeper can commit to any $(\alpha_R, \alpha_{\neg R})$ that keeps the Requesters indifferent. The continuum of equilibria remains. Commitment does not eliminate the multiplicity at the critical threshold. The detailed proof is left to the Appendix.

[3] At $p = q$, can standard refinements like the Intuitive Criterion eliminate any equilibria? Since all parties are indifferent, no off-equilibrium signal changes payoffs significantly. The Requester cannot send a costly signal to shift beliefs beneficially. The Gatekeeper's best response is linear in $P(I|a)$. Thus, refinement arguments do not rule out the continuum of equilibria. The Catch-22 nature remains robust.

Finally, it is useful to remark upon the appropriateness of a model that assumes one type of player might be Insane. Note that nothing will change in terms of the results if Insane types were non-strategic and essentially never choose to request not to fly because they want to be able to fly. In this case, while the Gatekeeper would infer that all the Insane never request not to fly (that is, $\beta_I = 1$, it is still optimal for the Gatekeeper to deny all requests to not fly so long as p is sufficiently low as characterised in Proposition 1. Thus, involving potentially irrational, Insane players does not change the outcomes thus far. Nonetheless, a solid takeaway from this analysis is that in high-stakes settings where sanity matters, investment in proper mental health evaluation is worthwhile.[4]

4. LABOUR INEXPERIENCE AS A CATCH-22

The previous sections presented a formal model of a Catch-22 situation, focusing on the logical paradox where access to a desired state (option A) is only granted to those who do not request it, thereby making it impossible to achieve for those who desire it. It was shown there that the equilibrium outcome, while observationally equivalent to a Catch-22 situation, was not supported by the inference associated with a Catch-22 situation. We now adapt this framework to a labour market setting, where the "experience" required by employers before hiring for a good job cannot be obtained without first being hired. This potentially creates a Catch-22 scenario: "You can't get a job without experience, but you can't get experience without a job."[5] Note, however, that this Catch-22 situation is simply a rule and does not have an underlying inference component. Thus, we explore the conditions under which the rule emerges in equilibrium.

In this labour market context, consider a fixed pool of workers of total mass N. A proportion p of these workers is inherently of high potential (H-

---

4 Unlike apparently authoring papers like this where surely one might reasonably question the sanity of the author in producing this research.

5 This thesis is exemplified by the work of Wilson (1987) and Wilson (1996). Wilson examined how the decline of manufacturing jobs and the rise of service-oriented economies led to a scarcity of employment opportunities in urban areas, particularly affecting African American communities. This scarcity created a cycle where individuals could not gain work experience due to the lack of available jobs yet were required to have experience to secure employment. Wilson argued that this cycle contributes to persistent poverty and social isolation in inner-city neighbourhoods.

type), while the remaining $(1-p)$ are low potential (L-type). Workers start as inexperienced, and no external signalling or credentialing is available. The only way for a worker to become "experienced" (and thus reveal or upgrade to their full productivity potential) is through employment. After one period of employment, an inexperienced worker's true type is revealed, or their productivity improves if the model assumes automatic upgrading. Either way, "experience" here means having been employed for a period, allowing the firm to identify or realise the worker's higher productivity if they are H-type.

The firm faces a downward-sloping labour demand function. Let E denote total employment. The marginal product of an experienced H-type worker is $H(E)$, and that of an experienced L-type worker is $L(E)$, with $H'(E) < 0$ and $L'(E) < 0$. Initially, workers are all inexperienced, so when the firm considers hiring from the pool of unknown workers, it must pay a wage equal to their expected productivity:

$$w_{\text{inexp}}(E) = pH(E) + (1-p)L(E).$$

If $pH(E) + (1-p)L(E) \leq 0$, the expected contribution of hiring another inexperienced worker is non-positive. In such a scenario, the firm might refuse to hire altogether. Without hiring, no worker gains experience, and the Catch-22 situation materializes: no one can become experienced (and thus no one can prove their higher productivity) because no one is hired. The equilibrium outcome depends critically on the choices made by employers. If employers are willing to take a risk and hire despite low expected productivity, some workers can gain experience and reveal their true productivity over time. If not, the economy remains in a no-hire equilibrium, trapping workers in inexperience forever.

When $pH(E) + (1-p)L(E) \leq 0$ from the start, no hiring occurs, leading to a full-blown Catch-22 with all workers trapped in inexperience. Now, assume conditions are more favourable at the outset so that at $E = 0$:

$$pH(0) + (1-p)L(0) > 0.$$

This means that initially, hiring an inexperienced worker is profitable. As a result, the firm begins by hiring some workers, granting them experience and revealing (or upgrading) their potential.

However, both $H(E)$ and $L(E)$ are downward-sloping in E, reflecting diminishing marginal returns as employment grows. As the firm adds more workers, the marginal productivity of both experienced and inexperienced hires declines. Eventually, the firm may reach an employment level E^* where:

$$pH(E^*)+(1-p)L(E^*)=0.$$

At this point, hiring another inexperienced worker yields zero expected productivity, leaving the firm indifferent or unwilling to hire further. Workers beyond E^* remain unemployed and inexperienced. They are stuck in a Catch-22: they need to be hired to gain experience and increase their productivity to $H(E)$ if they are H-type, but the firm will not hire them because doing so no longer yields positive returns.

In this scenario, the Catch-22 manifests as an equilibrium outcome affecting some but not all workers. Not all workers are blocked from employment; instead, a subset E^* is hired and becomes experienced, while the remainder $N-E^*$ is left in a state of permanent inexperience and unemployment. Those remaining workers face a Catch-22:

- They cannot become experienced without being hired.
- The firm will not hire them because at equilibrium $pH(E^*)+(1-p)L(E^*)=0$, hiring another inexperienced worker yields no net benefit.

This equilibrium differs from the earlier scenario, where no one was hired at all. Here, some are lifted out of inexperience, but the rest remain trapped. The firm's initial willingness to hire disappears once it saturates employment at E^*, where returns from additional inexperienced hires vanish.

These results are summarised in the following:

Proposition 2. *Suppose $pH(0)+(1-p)L(0)>0$ and that $H(E)$, $L(E)$ are strictly decreasing with E, with $H(E)>L(E)$ for all E. Then there exists an equilibrium employment level E^* such that:*

$$pH(E^*)+(1-p)L(E^*)=0,$$

and the firm hires exactly E^* workers, leaving $N-E^*$ workers permanently unemployed and inexperienced. The unemployed workers are stuck in a Catch-22, as they cannot gain experience without employment, and the firm will not hire them at zero expected marginal productivity.

Unlike the initial scenario where no one was hired (full Catch-22), here partial sorting occurs. A fraction E^* is extracted from the pool and becomes experienced, but the equilibrium stops short of employing everyone. The remaining $N - E^*$ are effectively locked out, unable to break into employment and upgrade their productivity. This final outcome still represents a Catch-22 equilibrium for the subset of workers left behind. They cannot become experienced because they cannot get hired, and the firm's equilibrium decision prevents any further expansion. While this might be a Catch-22 situation, it is basically a common result in labour economics; see, for example, Rosen (1981) and MacDonald (1982).

5. CONCLUSION

This paper has explored the logic, structure, and conditions of Catch-22 situations in a game-theoretic context. By examining two distinct yet parallel scenarios—a stylized version inspired by Joseph Heller's original Catch-22 and a labour market setting where workers cannot gain experience without first having a job—we have shown how such paradoxical outcomes can emerge as stable equilibria in interactive decision-making environments.

In both cases, a single underlying pattern emerges: individuals (Seekers) desire access to some valuable state or resource (e.g., relief from duties, a high-paying job), but the conditions imposed by an authority or gatekeeper (e.g., an officer, an employer) create a logical or strategic deadlock. The Seeker must possess a certain attribute or meet a prerequisite that can only be acquired after obtaining the very outcome they are seeking. As a result, the Seeker is trapped in a loop where attempting to become eligible invalidates their eligibility or where the inability to initiate a required change prevents them from ever qualifying.

Our models highlight that these Catch-22 outcomes do not arise from mere logical trickery but rather from the equilibrium interplay of strategies, incentives, and information constraints. In the literary scenario, a soldier's rational request to avoid a dangerous mission signals sanity, thereby disqualifying him from relief. In the labour market scenario, inexperienced workers remain unemployable if initial conditions fail to justify hiring them, blocking their only avenue to gain the experience needed to become employable. In both, the Catch-22 emerges as a stable configuration of actions and beliefs: the gatekeeper's best response leaves the Seeker stuck, while no profitable deviation exists for the Seeker to escape the predicament.

Our analysis suggests a general structure for identifying and characterizing Catch-22 situations in strategic settings:

1. Players: There are at least two types of players—Seekers who desire a beneficial outcome and Gatekeepers who control access to that outcome. The Seeker cannot autonomously achieve the desired state; the Gatekeeper must grant it or at least provide the necessary conditions.

2. Actions and Information: The Seeker takes actions (e.g., requesting relief, applying for a job) to demonstrate eligibility, while the Gatekeeper responds by accepting or rejecting based on observed signals, beliefs, and payoff structures. Typically, the Gatekeeper's decision rests on information that can only be credibly revealed or produced if the Seeker already had the outcome. This creates a circular dependency.

3. Conditions for a Catch-22: A Catch-22 arises under certain equilibrium conditions:

   - The Gatekeeper sets a condition that is logically or endogenously impossible for the Seeker to meet without already having what they seek.
   - There is no off-equilibrium path or profitable deviation that could break the cycle. Parameters (such as probabilities, costs, marginal products, or payoff differentials) push the equilibrium toward a no-improvement state for the Seeker.
   - Asymmetric information or incomplete verifiability ensures the Seeker cannot credibly demonstrate their eligibility without being granted the outcome first.

4. Why and When They Arise: Such paradoxes often occur in environments with:

   - Asymmetric Information: The Gatekeeper cannot directly observe the Seeker's true state, leading to screening rules that inadvertently exclude everyone.
   - Strict Prerequisites: The qualification the Gatekeeper demands is locked behind the very gate they guard.

- Unfavourable Parameter Ranges: The model's parameters (e.g., costs, payoffs, productivity levels) may fail to sustain any equilibrium in which the Seeker escapes the loop. In the labour market model, for instance, if hiring inexperienced workers is never profitable beyond a certain point, a nontrivial fraction of workers remains inexperienced forever.

The two scenarios we presented illustrate how different settings can produce structurally similar Catch-22 outcomes. In the original Catch-22 scenario, the condition "you must be insane to be relieved" and the logic that "requesting relief proves sanity" ensures no sane individual can qualify, trapping them in perpetual duty. In the labour market scenario, needing the experience to be hired and needing to be hired to gain experience creates an analogous loop. Initially, positive conditions may allow some workers to escape the trap, but diminishing returns and equilibrium constraints can still leave a subset of workers permanently locked out. In both cases, the paradox does not result from whimsical or ad hoc rules; rather, it emerges naturally from rational decision-making and equilibrium reasoning. The Gatekeeper chooses strategies that maximize their payoff given their beliefs, while the Seeker optimizes their actions within the constraints set by the Gatekeeper's equilibrium strategies. The paradoxical state is stable precisely because it aligns with all players' equilibrium incentives and the structural constraints of the game.

Understanding Catch-22 situations in formal terms opens avenues for both theoretical refinement and practical insight. Identifying the parameters or policy changes that could eliminate or mitigate the Catch-22 might be of direct interest. In the labour market case, for example, subsidies for hiring inexperienced workers or improved screening technologies that lower the cost of identifying high-potential workers could shift parameters and break the cycle. Similarly, in other domains—bureaucratic processes, credit markets, or admission systems—recognizing the equilibrium logic behind a no-win scenario may inspire interventions that relax prerequisites or provide credible signals, thus enabling Seekers to escape the loop.

In short, the general structure for Catch-22 situations involves a closed loop where the Seeker cannot become qualified without first obtaining the outcome, and the Gatekeeper's best response prevents granting that outcome without prior qualification. They are equilibrium outcomes, not anomalies. Understanding their conditions and character reveals not only why they occur but also hints at how changing underlying parameters might restore a more efficient and equitable equilibrium.

By casting Catch-22s into a strategic and equilibrium-centric framework, we gain a deeper understanding of how such paradoxes arise and persist. They are not mere narrative curiosities; they represent stable outcomes embedded in the logic of incentives, information, and constraints. Recognizing these structural elements paves the way for addressing them: by altering rules, improving information flows, or adjusting payoffs, it may be possible to shift from a no-improvement Catch-22 equilibrium to one in which Seekers can genuinely achieve the outcome they desire. In doing so, the analysis offered here contributes not only to economic theory but also to the understanding of institutional and organisational design where Catch-22-like predicaments pervade.

Appendix A

A.1. Proof of Proposition 1

Step 1: The Gatekeeper's Indifference Condition.

The Gatekeeper is indifferent after seeing R if:

$$E[G|R] = E[D|R].$$

Since:

$$E[G|R] = P(I|R)g_I(A) + (1 - P(I|R))g_S(A),$$
$$E[D|R] = P(I|R)g_I(B) + (1 - P(I|R))g_S(B).$$

Set $E[G|R] = E[D|R]$:

$$P(I|R)g_I(A) + (1 - P(I|R))g_S(A) = P(I|R)g_I(B) + (1 - P(I|R))g_S(B).$$

Rearrange terms:

$$P(I|R)(g_I(A) - g_I(B)) + (1 - P(I|R))(g_S(A) - g_S(B)) = 0.$$

Substitute $\Delta_I = g_I(A) - g_I(B) > 0$, $\Delta_S = g_S(A) - g_S(B) < 0$:

$$P(I|R)\Delta_I + (1 - P(I|R))\Delta_S = 0.$$

Solve for $P(I|R)$:

$$P(I|R)(\Delta_I - \Delta_S) = -\Delta_S \implies P(I|R) = \frac{-\Delta_S}{\Delta_I - \Delta_S} = q.$$

A similar derivation shows $P(I|\neg R) = q$ for indifference after $\neg R$.

Step 2: The Requester Indifference Conditions.

For the Sane type:

$$r_S(R) = \alpha_R r_S(A) + (1-\alpha_R) r_S(B), \quad r_S(\neg R) = \alpha_{\neg R} r_S(A) + (1-\alpha_{\neg R}) r_S(B).$$

If $r_S(R) = r_S(\neg R)$:

$$\alpha_R r_S(A) + (1-\alpha_R) r_S(B) = \alpha_{\neg R} r_S(A) + (1-\alpha_{\neg R}) r_S(B).$$

Expand:

$$\alpha_R r_S(A) + r_S(B) - \alpha_R r_S(B) = \alpha_{\neg R} r_S(A) + r_S(B) - \alpha_{\neg R} r_S(B).$$

Subtract $r_S(B)$:

$$\alpha_R(r_S(A) - r_S(B)) = \alpha_{\neg R}(r_S(A) - r_S(B)).$$

Since $r_S(A) - r_S(B) > 0$, we have $\alpha_R = \alpha_{\neg R}$. Similarly, for Insane:

$$r_I(R) = \alpha_R r_I(A) + (1-\alpha_R) r_I(B), \quad r_I(\neg R) = \alpha_{\neg R} r_I(A) + (1-\alpha_{\neg R}) r_I(B).$$

$r_I(R) = r_I(\neg R)$ gives, using the same algebraic procedure, $\alpha_R = \alpha_{\neg R}$. Both conditions yield $\alpha_R = \alpha_{\neg R} = \alpha$.

Step 3: Solve Under Different Cases of p vs. q.

Case (i) $p < q$: If the Gatekeeper wants to maximise their payoff, consider a deviation from always denying. If the Gatekeeper tries $\alpha > 0$, since $p < q$, the marginal effect of granting increases the likelihood of a worse scenario. In fact, the expected payoff decreases with α. Thus optimal $\alpha = 0$. With always deny, Requesters get B regardless. Neither type can improve by switching messages, as the Gatekeeper never changes action. Thus, in equilibrium, the Gatekeeper always denies, $(\alpha_R, \alpha_{\neg R}) = (0,0)$, and (β_S, β_I) arbitrary.

Case (ii) $p > q$: By symmetry, now granting is beneficial. If the Gatekeeper mixes, any reduction in granting lowers their payoff. So $\alpha = 1$; or always grant. Requesters always get A, and so have no incentive to deviate. Thus, in equilibrium, $(\alpha_R, \alpha_{\neg R}) = (1,1)$, (β_S, β_I) arbitrary.

Case (iii) $p = q$: Set $p = q$. For the Gatekeeper to remain indifferent, we need $P(I|R) = p$ and $P(I|\neg R) = p$. From $P(I|R) = \frac{p\beta_I}{p\beta_I + (1-p)\beta_S} = p$, solve:

$$p = \frac{p\beta_I}{p\beta_I + (1-p)\beta_S} \implies p\beta_I + (1-p)\beta_S = p\beta_I/p \implies \beta_S = \beta_I.$$

Similarly, from $P(I|\neg R)$, we get no contradiction; $\beta_S = \beta_I$ suffices.

With $\beta_S = \beta_I$, $P(I|R) = P(I|\neg R) = p$, and the Gatekeeper is indifferent. Any α works. Requesters are indifferent to messages. Thus, there is a continuum of equilibria:

$$\beta_S = \beta_I \in [0,1], \quad \alpha_R = \alpha_{\neg R} = \alpha \in [0,1].$$

This proves the proposition.

A.2. Proof of Equilibria When the Gatekeeper Commits First

In this subsection, we consider the alternative timing where the Gatekeeper announces and commits to a policy before the Requester chooses their message. Specifically, the timing is:

1. Nature determines the Requester's type: S with probability $1 - p$ or I with probability p.

2. The Gatekeeper publicly commits to a policy $(\alpha_R, \alpha_{\neg R})$, where α_R is the probability of granting after observing R, and $\alpha_{\neg R}$ is the probability of granting after observing $\neg R$.

3. After seeing the Gatekeeper's commitment, the Requester chooses R or $\neg R$ to maximize their own expected payoff given their type.

We consider a Subgame Perfect Equilibrium (SPE). The Gatekeeper chooses $(\alpha_R, \alpha_{\neg R})$ first. Given that policy, the Requester, knowing their type, chooses R or $\neg R$. Because the Requester moves last, they pick the action that maximizes their expected utility, taking $(\alpha_R, \alpha_{\neg R})$ as given. The Gatekeeper anticipates this response and chooses an optimal policy.

The Requester's Best Response Given $(\alpha_R, \alpha_{\neg R})$

For a given Gatekeeper policy: - If the Requester is Sane, their expected payoffs are:

$$r_S(R) = \alpha_R r_S(A) + (1-\alpha_R) r_S(B), \quad r_S(\neg R) = \alpha_{\neg R} r_S(A) + (1-\alpha_{\neg R}) r_S(B).$$

Since $r_S(A) > r_S(B)$, the Sane type compares $r_S(R)$ and $r_S(\neg R)$:

$$r_S(R) - r_S(\neg R) = (\alpha_R - \alpha_{\neg R})(r_S(A) - r_S(B)).$$

- If the Requester is Insane:

$$r_I(R) = \alpha_R r_I(A) + (1-\alpha_R) r_I(B), \quad r_I(\neg R) = \alpha_{\neg R} r_I(A) + (1-\alpha_{\neg R}) r_I(B).$$

With $r_I(B) > r_I(A)$, we have:

$$r_I(R) - r_I(\neg R) = (\alpha_R - \alpha_{\neg R})(r_I(A) - r_I(B)).$$

Note $(r_I(A) - r_I(B)) < 0$, so if $\alpha_R > \alpha_{\neg R}$, the Insane type is more likely to get A (unwanted), decreasing their utility relative to $\neg R$; if $\alpha_R < \alpha_{\neg R}$, the reverse is true.

The Gatekeeper's Anticipation

The Gatekeeper knows that after commitment, the Requester chooses the message to maximize their payoff. Consider what happens under different configurations of $(\alpha_R, \alpha_{\neg R})$:

1. If $\alpha_R > \alpha_{\neg R}$: the Sane type prefers R since that yields a higher probability of A, their preferred outcome. The Insane type prefers $\neg R$ since $\alpha_R > \alpha_{\neg R}$ makes R more likely to yield A, which they dislike. Thus, the Insane type avoids R. In this scenario, the messages separate types: Sane chooses R, Insane chooses $\neg R$.

2. If $\alpha_R < \alpha_{\neg R}$ the Sane type prefers $\neg R$ (since $\neg R$ has a higher probability of A now). The Insane type prefers R (since R now has a lower probability of A and thus a higher chance of B, their preferred outcome). Here, the messages also separate types but in the opposite manner: Sane chooses $\neg R$, Insane chooses R.

3. If $\alpha_R = \alpha_{\neg R}$ both Sane and Insane are indifferent between R and $\neg R$. Hence, the Requesters may mix arbitrarily. There is no separating pattern; either type could choose R or $\neg R$ with any probability.

Given this, the Gatekeeper's payoff depends on how the actions map to outcomes for each type. We use the same logic as in the main text. The Gatekeeper would like Sane types to end up with B and Insane types to end up with A.

Now, consider the three main cases regarding p and q:

Case 1: $p < q$**.**

If $p < q$, recall from the main text analysis that when the Gatekeeper attempts to equalize or mix, the equilibrium in the earlier scenario leads to always denying. Here, because the Gatekeeper moves first, can they do better? Suppose the Gatekeeper tries some $\alpha > 0$. If α_R and $\alpha_{\neg R}$ differ, the Requester separates: one type gets more A, the other gets more B. The Gatekeeper wants Insane to get A and Sane to get B, but if $p < q$, the population is more likely Sane. If the Gatekeeper chooses $\alpha_R > \alpha_{\neg R}$, Sane chooses R, gets mostly A (which the Gatekeeper does not want). If $\alpha_R < \alpha_{\neg R}$, Sane chooses $\neg R$, gets mostly A anyway (just switching messages), still against the Gatekeeper's preference.

By trying to induce a difference between α_R and $\alpha_{\neg R}$, the Gatekeeper inadvertently ensures that the Sane type will gravitate to the message that yields A. Since $p < q$, the Gatekeeper's expected payoff is reduced by granting frequently because the likely Sane Requester gets their preferred outcome A, which the Gatekeeper dislikes.

To avoid this, the Gatekeeper sets $\alpha_R = \alpha_{\neg R} = 0$. This yields no incentive for Sane or Insane to prefer one message over the other; both get B. Insane is happy with B, Sane is not, but cannot do better by switching. The Gatekeeper's expected payoff is:

$$\Pi_{\text{deny}} = p g_I(B) + (1-p) g_S(B).$$

This is exactly the same outcome as in the previously analyzed model where the Requester moved first. The Gatekeeper cannot improve upon always denying. Hence, if $p < q$, the Gatekeeper's optimal committed policy is $(\alpha_R, \alpha_{\neg R}) = (0,0)$.

Case 2: $p > q$**.**

By symmetry, if the probability of Insanity is high, the Gatekeeper prefers to ensure Insane types get A. If the Gatekeeper tries $\alpha_R < \alpha_{\neg R}$ or vice versa, they end up granting a scenario where Sane might get A. With $p > q$, there are relatively more Insane types, so to ensure high expected payoff, the Gatekeeper sets $\alpha_R = \alpha_{\neg R} = 1$, always granting. In this scenario, both types get A. Insane prefers B, but cannot force the Gatekeeper to deny. Sane is satisfied with A, Insane cannot do better. The Gatekeeper's payoff is:

$$\Pi_{\text{grant}} = pg_I(A) + (1-p)g_S(A).$$

Since $p > q$, this aligns with the Gatekeeper's preference to favor outcomes suitable for Insane. No mixing improves the Gatekeeper's payoff beyond always granting.

Case 3: $p = q$**.**

If $p = q$, the Gatekeeper is exactly indifferent when facing balanced posteriors. By choosing $\alpha_R = \alpha_{\neg R} = \alpha$ for any $\alpha \in [0,1]$, the Requesters remain indifferent. They can mix in any proportion. No profitable deviation exists for the Gatekeeper by breaking this balance, as any asymmetry would allow the Requesters to self-select into messages that yield an undesired pairing of types and outcomes. Hence, a continuum of equilibria emerges:

$$(\alpha_R, \alpha_{\neg R}) = (\alpha, \alpha), \quad \alpha \in [0,1],$$

with the Requesters mixing arbitrarily. This exactly mirrors the multiplicity found in the original timing scenario. The Gatekeeper, anticipating the Requesters' best responses, cannot improve by picking a unique pure strategy other than a corner solution, but since $p = q$, corner solutions are no better than mixing solutions.

Conclusion for the Gatekeeper-First Scenario

When the Gatekeeper commits first:

- If $p < q$, the Gatekeeper chooses $(\alpha_R, \alpha_{\neg R}) = (0,0)$, always denying. Requesters mix arbitrarily, equilibrium outcome matches the previously found pure equilibrium.

- If $p > q$, the Gatekeeper chooses $(\alpha_R, \alpha_{\neg R}) = (1,1)$, always granting. Requesters again mix arbitrarily, equilibrium outcome matches the pure equilibrium of always granting.

- If $p = q$, the Gatekeeper can choose $(\alpha_R, \alpha_{\neg R}) = (\alpha, \alpha)$ for any $\alpha \in [0,1]$. Requesters mix freely, resulting in a continuum of equilibria that parallel the original scenario.

Thus, committing early does not remove the multiplicity at $p = q$, nor does it lead to a better outcome for the Gatekeeper when $p \neq q$. The results are fully consistent with the analysis conducted under the original timing.

A.3. Proof of Proposition 2

Step 1: Existence of E^*. Start from $E = 0$ where $pH(0) + (1-p)L(0) > 0$. By continuity and the fact that $H(E)$ and $L(E)$ decrease with E, there must be some largest $E > 0$ at which $pH(E) + (1-p)L(E) = 0$. Define E^* as that point:

$$pH(E^*) + (1-p)L(E^*) = 0.$$

Such an E^* exists because at $E = 0$ the value is positive, and as $E \to \infty$, $H(E), L(E)$ become very low (possibly negative or approaching zero), ensuring that eventually the expected productivity crosses zero.

Step 2: Optimality of Hiring Up to E^***.** For $E < E^*$, we have $pH(E) + (1-p)L(E) > 0$, so hiring another inexperienced worker is profitable. The firm continues hiring until E reaches E^*. At E^*, the next marginal hire yields zero expected return, so the firm stops hiring further to avoid negative gains.

Step 3: Catch-22 for $N - E^*$ **Workers.** At equilibrium E^*, there are $N - E^*$ workers who remain untested and inexperienced. Their potential productivity could be high if they are H-type, but they never gain the chance to demonstrate this because the firm refuses to hire more workers once the marginal expected productivity hits zero. These workers are stuck in the situation that defines Catch-22: (i) they need a job to gain experience and reveal or achieve higher productivity and (ii) the firm will not hire them now that the expected return is zero at the margin.

References

Brams, S. J. (1999). To mobilize or not to mobilize: Catch-22s in international crises. *International Studies Quarterly*, *43*(4), 621–640.

Brams, S. J., & Jones, C. B. (1999). Catch-22 and king-of-the-mountain games: Cycling, frustration, and power. *Rationality and Society*, *11*(2), 139–167.

Carmichael, F. (2002). Catch-22 in a signalling game. *Journal of Institutional and Theoretical Economics*, *158*(3), 375–392.

Colgan, C. S., & Quinlin, G. (1997). The catch-22 of congestion pricing. *Transportation Quarterly*, *51*(2), 117–133.

Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, *50*(6), 1431–1451.

de Rouffignac, A. (2000). Supply vs. demand: Catch-22 for us gas. *Energy Economist*(219), 6–9.

Flood, P. C., & Toner, B. (1997). La catch 22? *British Journal of Industrial Relations*, *35*(1), 123–124.

Goldstein, L. (2004). The barber, russell's paradox, catch-22, god, contradiction, and more. In G. Priest, J. C. Beall, & B. P. Armour-Garb (Eds.), *The Law of Non-Contradiction: New Philosophical Essays* (pp. 295–313). New York: Oxford University Press.

Heller, J. (1961). *Catch-22*. New York: Simon & Schuster.

MacDonald, G. M. (1982). Information in production. *Econometrica*, *50*(5), 1143–1162.

Rosen, S. (1981). An equilibrium approach to the labor market. *Econometrica*, *49*(3), 565-584.

Stein, M. K. (1999). The catch-22 of retirement. *Journal of Financial Planning*, *12*, 28–30.

Wilson, W. J. (1987). *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy*. University of Chicago Press.

Wilson, W. J. (1996). *When Work Disappears: The World of the New Urban Poor*. Vintage Books.

Journal of Mechanism and Institution Design
ISSN: 2399-844X (Print), 2399-8458 (Online)
DOI:10.22574/jmid.2025.12.003

POLYNOMIALLY TRACTABLE CASES IN THE POPULAR ROOMMATES PROBLEM

Erika Bérczi-Kovács, Eötvös Loránd University, Hungary
erika.berczi-kovacs@ttk.elte.hu

Ágnes Cseh, University of Bayreuth, Germany
agnes.cseh@uni-bayreuth.de

Kata Kosztolányi, Eötvös Loránd University, Hungary
kata.kosztolanyi98@gmail.com

Attila Mályusz, University of Freiburg, Germany
atimaly@cs.uni-freiburg.de

ABSTRACT

The input of the popular roommates problem consists of a graph $G = (V, E)$ and for each vertex $v \in V$, strict preferences over the neighbors of v. Matching M is more popular than M' if the number of vertices preferring M to M' is larger than the number of vertices preferring M' to M. A matching M is called *popular* if there is no matching M' that is more popular than M. Faenza et al. (2019) and Gupta et al. (2021) proved that determining the existence of a popular matching in a popular roommates instance is NP-complete. In this paper we identify a class of instances that admit a polynomial-time algorithm for the problem. We also test these theoretical findings on randomly generated instances to determine the existence probability of a popular matching in them.

Keywords: Popular matching, stable matching, complexity.

JEL Classification Numbers: C78, C63.

E. Bérczi-Kovács was supported by the HUN-REN-ELTE Egerváry Research Group on Combinatorial Optimization, by the János Bolyai Research Scholarship of the Hungarian Academy of Science and by the project no. FK 132524, provided by the National Research, Development and Innovation Fund of Hungary. We declare that there are no conflicts of interest. Any errors are our own.

1. INTRODUCTION

OUR input is an instance of the roommates problem, consisting of a graph $G = (V, E)$, where each vertex is equipped with a strict preference list over its neighbors. A matching M is *stable* if there is no *blocking pair* with respect to M, in other words, there is no pair of vertices (a, b) such that a is either unmatched or prefers b to $M(a)$ (a's partner in M) and similarly, b is either unmatched or prefers a to $M(b)$. Irving (1985) gave a polynomial-time algorithm to decide whether such an instance admits a stable matching.

In this paper we consider *popularity*, a notion that is more relaxed than stability. For a vertex $v \in V$, v's preference list over its neighbors can be extended naturally to preferences over matchings as follows: v prefers matching M to matching M' if (i) v is matched in M and unmatched in M' or (ii) v is matched in both and prefers $M(v)$ to $M'(v)$.

The motivation of popularity comes from voting. Suppose an election is held between M and M' where each vertex casts a vote for the matching it prefers. We call a matching M popular in the instance if it never loses an election to another matching M'. In voting terminology, each popular matching is a weak *Condorcet winner* (Condorcet, 1785) in the corresponding voting instance.

Stable matchings are always popular (Chung, 2000). Thus, if an instance admits a stable matching, then the existence of a popular matching is also guaranteed, and it can be found using Irving's algorithm for finding a stable matching (Irving, 1985). Some instances of the stable roommates problem do not admit a stable solution, yet they admit a popular matching, as demonstrated by Figure 1, first presented by Biró et al. (2010).

Validating whether a given matching is popular can be done in polynomial time (Biró et al., 2010). Recently, Faenza et al. (2019) and Gupta et al. (2021) solved a well-known open problem on the complexity of deciding whether a popular matching exists in a popular roommates instance and showed that the problem is NP-complete (Biró et al., 2010; Cseh, 2017; Huang & Kavitha, 2013a, 2021; Manlove, 2013). This hardness extends to instances with complete preference lists (Cseh & Kavitha, 2021). However, this latter hardness result is only valid for instances on complete graphs with an *even* number of vertices. It is known that when n, the number of vertices in G is *odd*, a matching in an instance on a complete graph G on n vertices is popular if and only if it is stable (Cseh & Kavitha, 2021). Since Irving's algorithm (Irving, 1985) can decide if

$$\begin{array}{llll} a: & b, & d, & e \\ b: & d, & a, & e \\ d: & a, & b, & e \\ e: & d, & b, & a \end{array}$$

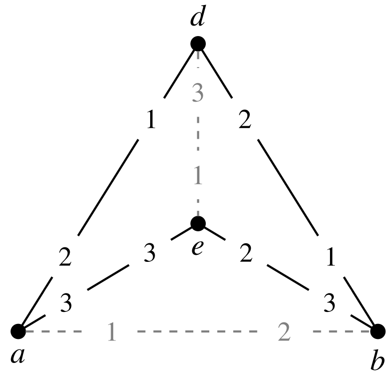

Figure 1: The preferences of each vertex are shown on the left hand-side and also as labels on the edges in the graph. The dashed gray edges mark the unique popular matching $M = \{(a, b), (d, e)\}$. It is blocked by the edge (b, d). The instance admits no stable matching.

an instance admits a stable matching or not, the popular roommates problem in an instance on a complete graph G can be efficiently solved for odd n. It is easy to see that the proof in Cseh & Kavitha (2021) extends to graphs with minimum degree $n - 2$ as well: even in those instances, a matching is popular if and only if it is stable.

1.1. Our contribution

In this paper, we stretch this observation even further and show that for instances on *high-degree* graphs with an *odd n*, there is a polynomial-time algorithm that solves the popular roommates problem. More precisely, our algorithm runs in polynomial time for a constant c, where $n-c$ is the minimum degree in the graph. The merit of our result mainly lies in the deep investigation of the unusual behaviour for a natural decision problem in combinatorial optimization, namely to be efficiently solvable when n has one parity and become NP-complete when n has the other parity. A clear difference between the two cases for matching problems is that for an odd n, at least one vertex remains unmatched in any matching. We capitalize on this simple observation as follows.

Our key technical result is an algorithm to decide whether there exists a popular matching that leaves exactly a given non-empty set of vertices uncovered. We prove that popular matchings with this property can be decomposed into a stable and a popular part. We iterate through the possible popular parts and construct the fitting stable part in a carefully designed subroutine. More precisely, we establish the following results.

- Theorem 2: For a vertex set U, a popular matching that leaves exactly the vertices in U uncovered or a proof for its non-existence can be calculated in $O\left(n^{|Z|} \cdot (m + |Z|^3)\right)$ time, where Z is the set of vertices that are neither in U nor adjacent to any vertex in U.
- By iterating through all possible sets of uncovered vertices, our algorithm is able to decide whether a popular matching exists in $O\left(n^c \cdot (m + c^3)\right)$ time in instances with an odd n and $\deg(v) \geq n - c$ for all $v \in V$. It also delivers a maximum size popular matching if the iteration happens in an increasing order of $|U|$.
- For instances with an even n, our algorithm is only able to decide whether a non-perfect popular matching exists. This is not surprising, because deciding whether a (perfect) popular matching exists in an instance on a complete graph with an even n is NP-complete (Cseh & Kavitha, 2021).

We also test our algorithm on randomly generated Erdős–Rényi graphs (Erdős & Rényi, 1960) and conclude that on instances with a high minimum degree, our method is significantly faster than checking each matching in the instance for popularity. Furthermore, we also tally the instances that admit a popular matching, and within this set, the instances that admit no stable matching. Ours is the first experimental study on the existence probability of popular matchings in randomly generated non-bipartite instances.

1.2. Literature review

Bipartite instances. The notion of popularity was first introduced in 1975 by Gärdenfors (1975), who showed that popular matchings always exist in bipartite instances, since stable matchings always exist there (Gale & Shapley, 1962) and every stable matching is popular. In bipartite graphs, all stable matchings have the same size, at least $\frac{1}{2}|M_{\max}|$ where $M_{\max}$ is a maximum-size matching. Stable matchings are minimum-size popular matchings (Gärdenfors, 1975; Huang & Kavitha, 2013b), and maximum-size popular matchings have size at least $\frac{2}{3}|M_{\max}|$ (Kavitha, 2014), and all of them cover the same set of vertices (Hirakawa et al., 2015; Brandl & Kavitha, 2019). Faenza et al. (2019) showed that it is NP-hard to decide if an instance admits a popular matching that is neither a min-size nor a max-size popular matching. Not only the size of popular matchings, but also the set of covered vertices has received interest. It follows from the proof of Lemma 5 in Cseh et al. (2022) that in the bipartite case, it is NP-complete to decide whether a popular matching that leaves some vertices (and possibly others) unmatched exists. Our algorithm on finding a popular matching that leaves exactly a given set of vertices unmatched in a special graph class complements this negative result.

Hardness for non-bipartite instances. The proof that every stable matching is popular holds in non-bipartite instances as well (Chung, 2000). Huang & Kavitha (2021) showed that the polytope of popular fractional matchings is half-integral in the non-bipartite case. This means that for an arbitrary rational-valued weight function on the edges, one can compute a maximum weight popular half-integral matching in polynomial time. They also showed that the problem of computing an integral maximum weight popular matching in a non-bipartite instance is NP-hard. Finding a maximum cardinality popular matching in instances admitting a popular matching has also been shown to be NP-hard by Brandl & Kavitha (2019). The breakthrough on the complexity of the popular roommates problem was made by two independent NP-completeness proofs given by Faenza et al. (2019) and Gupta et al. (2021) in 2019, respectively.

Tractable non-bipartite cases. Most positive results in the setting are also very recent. The only known tractable subclasses of popular matchings are the class of stable matchings and the class of so-called 'strongly dominant matchings', which is a subclass of max-size popular matchings (Faenza et al., 2019). If the underlying graph G has a bounded treewidth, then the min-cost popular matching problem can be solved in polynomial time (Faenza et al., 2019). Kavitha (2024) provided a simply exponential time algorithm for the popular roommates problem. Interestingly, its running time was slightly decreased very recently by Palmer & Pálvölgyi (2022), who improved the upper bound on the number of stable matchings in bipartite matching instances. However, her algorithm stays exponential for the special class of graphs we discuss in this paper. Huang & Kavitha (2013a) proved that each roommates instance admits a matching that has the approximability measure called 'unpopularity factor' of $O(\log n)$.

Experimental existence results. The existence probability of a stable matching in randomly generated roommates instances has been an actively researched topic for decades, which lead to both theoretical (Irving, 1985; Pittel, 1993; Pittel & Irving, 1994) and experimental findings (Prosser, 2014; Mertens, 2015; Erdem et al., 2020). Until now, randomly generated popular matching instances have only been studied in the context of bipartite graphs, both with one-sided (Mahdian, 2006; Itoh & Watanabe, 2010; Ruangwises & Itoh, 2019) and two-sided preferences (Abraham et al., 2007).

2. PRELIMINARIES

Consider a graph $G = (V, E)$ with n vertices and m edges. Let $N(u)$ denote the vertices adjacent to $u \in V$ in G, and let $N(U)$ for $U \subseteq V$ be the set of vertices that are adjacent to at least one vertex in U. Furthermore, let $G[U]$ be the graph spanned by the vertex set U.

Each vertex ranks all adjacent vertices in a strict order of preference. We write $v \succ_u w$ if u prefers v to w. A *matching* $M \subseteq E$ is a set of edges such that each vertex is incident to at most one edge in M. For convenience, we denote u's partner in M by $M(u)$, and write $M(u) = u$ if u is unmatched in M. We assume that $v \succ_u u$ for all $v \in N(u)$. In other words, each vertex prefers being matched along any of its edges to staying unmatched.

We now introduce the edge labeling technique that is standard in the field (Huang & Kavitha, 2013b). Let M be a matching in the instance. For each edge $(u, v) \notin M$, we define the vote of vertex u between the edge (u, v) and M as follows:

$$\mathrm{vote}_u(v, M) = \begin{cases} + & \text{if } v \succ_u M(u); \\ - & \text{if } M(u) \succ_u v. \end{cases}$$

The next step is to label every edge $(u, v) \notin M$ by the pair $(\mathrm{vote}_u(v, M), \mathrm{vote}_v(u, M))$. Thus every non-matching edge has a label in $\{(\pm, \pm)\}$; in the example in Figure 1 these labels are $(-, +), (+, +), (-, -), (-, -)$ for the edges $(a, d), (d, b), (b, e), (e, a)$ outside of the matching $M = \{(a, b), (d, e)\}$. Note that an edge is labeled $(+, +)$ if and only if it blocks M. Let G_M be the subgraph of G obtained by deleting edges labeled $(-, -)$ from G. The following theorem characterizes popular matchings in G.

Theorem 1 (Huang & Kavitha (2013b)). *M is popular in G if and only if G_M does not contain an improving path or cycle, defined as any of the following with respect to M:*

1. *an alternating cycle with a $(+,+)$ edge;*
2. *an alternating path with two disjoint $(+,+)$ edges;*
3. *an alternating path with a $(+,+)$ edge and an unmatched vertex as an end vertex.*

Using the above characterization, it can be easily checked whether a given matching M is popular (Faenza et al., 2019): M is popular if and only if it is a maximum weight perfect matching in the weighted graph we get by first adding a loop to each vertex and then assigning weight -1 to those loops that belong to vertices covered by M, 2 to all $(+,+)$ edges, -2 to all $(-,-)$ edges, and 0 to all other edges and loops. However, this only settles the complexity of verification.

Structure of the paper. We first explain the high-level idea of our algorithm in Section 3. Its proof of correctness and the exact implementation details are then given in Section 4. Section 5 contains a detailed example of the execution. In Section 6 we analyze the performance of our algorithm on randomly generated instances. We discuss relations to other decompositions from the literature and pose open questions in Section 7. Two full example runs are to be found in the appendix.

3. OUR ALGORITHM

As already observed in Section 1, if the instance admits a stable matching, then that is popular as well, so the existence question is only interesting for instances that do not admit any stable matching. From now on, we restrict our attention to these instances. Let U be a non-empty set of vertices in such an instance. We now present an algorithm that checks whether there is a popular matching in G that leaves exactly U uncovered. Notice that the answer is trivially 'NO' if U spans any edge in G.

To a given U, let $Z = V \setminus N(U) \setminus U$ denote the set of those vertices that are not in U and also not adjacent to *any* vertex in U. It is clear that the set Z is well-defined for each set U. If there is a popular (but unstable) matching in the instance, we find it as the union of a popular matching covering Z and a stable matching covering the rest of $V \setminus U$. In our algorithm (see Algorithm 1 for a pseudocode), we build a set $\mathcal{P}_Z$ of initial matchings covering Z and test for each $P_Z \in \mathcal{P}_Z$ whether it can be completed to a popular (but unstable) matching that leaves exactly U uncovered. The construction of $\mathcal{P}_Z$ is simple: we list all matchings in which each edge has at least one end vertex in Z and also cover each vertex in Z. Then, we examine each such $P_Z \in \mathcal{P}_Z$ in two tests.

1. The first test checks if P_Z is a popular but unstable matching in $G' = G[Z \cup P_Z(Z) \cup U]$. If not, we stop testing this P_Z. We remind the reader that G' is the graph spanned by the vertex set Z, the vertices matched to any vertex in Z in P_Z, and the (desired) set of unmatched vertices U.
   - If P_Z is not popular in G', then P_Z cannot be extended to a desired popular matching in G, as shown in Claim 1.
   - A similar statement applies to the case when P_Z is (popular and) stable in G'. More precisely, we prove in Corollary 1 that edges blocking a popular matching are incident to vertices covered by P_Z.
2. If P_Z has passed the first test, then there is at least one alternating path traversing through a blocking edge in G'. In Claim 3 we show that matching P_Z can be extended to a popular matching in G with edge set S leaving exactly U uncovered if and only if S is a stable complete matching in a transformed instance. Later, in Section 4.3 we will show how the existence of a desired stable matching can be checked in polynomial time.

4. PROOF OF CORRECTNESS

In Sections 4.1 and 4.2 we prove the necessity of our two tests in lines 4 and 5 in our pseudocode. Then in Section 4.3 we elaborate on the implementation of the second test. Finally in Section 4.4 we complete the proof of correctness and provide a running time analysis.

4.1. First test

Our first claim shows the necessity of popularity in line 4. More precisely, we show that if a matching P_Z is not popular in G', then it cannot be extended to a

---

Algorithm 1: Routine checking if the independent set $U \neq \emptyset$ can be the set of uncovered vertices in a popular matching

---

Input: A popular roommates problem instance that admits no stable matching and a set of independent vertices $\emptyset \neq U \subseteq V$.

Output: A popular matching covering exactly $V \setminus U$ or a 'NO' answer.

1 $Z :=$ set of vertices in $V \setminus U$ that are not adjacent to any vertex in U

2 $\mathcal{P}_Z :=$ set of matchings P_Z covering Z such that for every edge $(p, q) \in P_Z$: $Z \cap \{p, q\} \neq \emptyset$

3 **for** $P_Z \in \mathcal{P}_Z$ **do**

4   **if** *P_Z is popular but unstable in $G' = G[Z \cup P_Z(Z) \cup U]$* **then**

5   **if** *a complete stable matching S exists on $N(U) \setminus P_Z(Z)$ fulfilling the conditions in Claim 3 for the existence of a popular matching P that contains P_Z and leaves exactly U uncovered* **then**

            return $P_Z \cup S$

6 **return** *'NO'*

---

popular matching in G, and thus we can proceed to testing the next candidate for P_Z.

Claim 1. *If the constructed matching P_Z is not popular in G', then no matching P containing P_Z is popular in G.*

Proof. As P_Z is not popular in G', the graph G'_{P_Z} admits an improving path or cycle, as shown in Theorem 1. This improving path or cycle also exists in G_P if $P_Z \subseteq P$, since G'_{P_Z} is a subgraph of G_P. □

Next we turn to proving the necessity of instability in line 4. The key observation is that edges blocking a popular matching P cannot occur anywhere in the graph, but only between vertices that P matches to Z. This observation is a corollary of the following claim.

Claim 2. *Let P be a popular matching that leaves exactly U uncovered. Vertex q is in Z if edge $(p, q) \in P$ can be reached on an alternating path ϱ in the graph G_P such that*

- *ϱ starts with an edge that blocks P and*

- *ϱ ends at vertex q.*

Proof. Suppose by contradiction that there is an edge $(p,q) \in P$ that can be reached on an alternating path in G_P from a blocking edge such that q is the last vertex on that path, yet $q \notin Z$. By the definition of Z, $q \notin Z$ means that $q \in U \cup N(U)$. Since no vertex in U is matched in P and $(p,q) \in P$, vertex q must be in $N(U)$. Now we take the alternating path in G_P that traverses (p,q). We lengthen this path past q toward U along the $(+,-)$ edge that connects q to some $u \in U$—notice that all edges have a + label at the unmatched vertex $u \in U$. With this new path we have shown that to P, there is an alternating path that traverses a blocking edge and ends in an unmatched vertex, which contradicts the assumption on the popularity of P due to Theorem 1. □

Corollary 1. *Let P be a popular matching that leaves exactly U uncovered. Each edge blocking P connects two vertices that P matches to vertices in Z.*

Proof. A blocking edge (p_1,p_2) and the matching edges $\{(q_1,p_1),(q_2,p_2)\} \subseteq P$ at its two end vertices build an alternating path. We have shown that q_1 and q_2 are in Z, because they can be reached on an alternating path from (p_1,p_2) that fulfills the conditions in Claim 2. □

Corollary 1 implies that if P_Z is stable in G', then it cannot be extended to the desired popular but unstable matching P in G.

4.2. Second test

The input of the second test is a popular roommates problem instance and an initial matching P_Z that has passed the first test. The output is a complete stable matching S on the set of vertices that are adjacent to at least one vertex in U and not covered by P_Z, or a 'NO' answer.

For better readability, we introduce the notation $V' = Z \cup P_Z(Z) \cup U$ for the vertex set of graph G'. Notice that $V \setminus V' = N(U) \setminus P_Z(Z)$, which is the set of vertices that are adjacent to at least one vertex in U and not covered by P_Z. Let us denote by D the set of 'dangerous' vertices. A vertex z is in D if it can be reached on an alternating path from a blocking edge in G'_{P_Z} such that counting from the blocking edge, z is the further end of the matching edge $(P(z),z)$ on the path (see Figure 2). Note that by applying Claim 2 to $P = P_Z$, we get that $D \subseteq Z$.

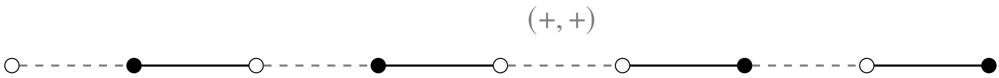

Figure 2: Edges in P_Z are solid, and the dashed edges are also in G'_{P_Z}, in particular, the blocking edge is marked by $(+,+)$. Vertices in D are denoted by black color.

Dangerous vertices play a crucial role in our second test in line 5. These vertices can be reached on an alternating path from a blocking edge, which path must be discontinued in G_P before it reaches U. Our next claim guarantees that we extend P_Z to a P that fulfills this criterion, provided it is possible for the given P_Z at all.

Claim 3. *Assume that P_Z is popular but unstable in G'. In G there is a popular matching P that contains P_Z and leaves exactly U uncovered if and only if $S := P \setminus P_Z$ is a stable matching in $G[V \setminus V']$, covers exactly the vertices in $V \setminus V'$, and induces the following edge labeling with respect to $P = P_Z \cup S$:*

(a) $(+,-)$ *for edges incident to U, with a $'+'$ at the vertex in U,*

(b) $(-,-)$ *for edges between D and $V \setminus V'$,*

(c) $(+,-)$ *for an edge (v', x) where $v' \in V' \setminus (D \cup U)$, $x \in V \setminus V'$, and v' prefers x to $P(v')$, with a $'+'$ at v'.*

Proof. Let us first assume that in G there is a popular matching P that contains P_Z and leaves exactly U uncovered. Since $V \setminus V' = N(U) \setminus P_Z(Z)$ and P_Z must cover all vertices in Z by construction, $S = P \setminus P_Z$ must cover exactly the vertices in $V \setminus V'$. We will first show that S is a stable matching in $G[V \setminus V']$, and then prove that P induces an edge labeling satisfying the three conditions above.

The stability of S is easy to show. We have seen in Corollary 1 that in a popular matching every blocking edge is adjacent to two edges in P_Z, hence $S = P \setminus P_Z$ is stable in $G[V \setminus V']$.

We now check the edge labeling property in each condition separately.

(a) Edges incident to U trivially have a $'+'$ at the vertex in U. The other label must be a $'-'$, otherwise we find a blocking edge at an unmatched vertex, which contradicts the popularity of P.

(b) Suppose by contradiction that there are vertices $v \in V \setminus V'$ and $z \in D$ such that (v, z) is not labelled $(-, -)$ with respect to P. By the definition of V', there exists a vertex $u \in U$ such that $\langle u - P(v) - v - z \rangle$ is an augmenting path in G_P, furthermore by the definition of D, a blocking edge is reachable from z via an alternating path in G_P. This latter path is disjoint from $(v, P(v))$, because a blocking edge can only occur in $G[Z \cup P_Z(Z)]$ (see Corollary 1) and from $z \in D$ it can only be reached via an alternating path that also runs in this subgraph, while both v and $P(v)$ are in $V \setminus V'$. The concatenation of these two paths is therefore an augmenting path to P in G_P, which contradicts the popularity of P.

(c) Since v' prefers x to $P(v')$, a $'+'$ at v' is guaranteed. Suppose by contradiction that there are vertices $v' \in V' \setminus (D \cup U)$ and $x \in V \setminus V'$ such that (v', x) blocks P. Then by the definition of $V \setminus V'$, there is a vertex $u \in U$ such that $\langle u - P(x) - x - v' - P(v') \rangle$ is an augmenting path in G_P, which contradicts the popularity of P.

We now turn to the opposite direction and its corresponding proof. Assume that S is stable in $G[V \setminus V']$, it covers exactly $V \setminus V'$, and the three conditions on edge labeling are fulfilled. Since $(Z \cup P_Z(Z)) \cup (V \setminus V') = V \setminus U$, matching $P = P_Z \cup S$ leaves exactly U uncovered. We will next utilize Theorem 1, the characterization of popular matchings by Huang & Kavitha (2013b), to show that P is popular in G.

1. G_P admits no alternating cycle that contains a blocking edge.
   Neither P_Z, nor S creates an alternating cycle with a blocking edge in the subgraphs G_{P_Z} on the vertex set $Z \cup P_Z(Z)$ and G_S on the vertex set $V \setminus V'$, respectively. Therefore, for such an alternating cycle in G_P, edges between the two sets of vertices must connect paths to form a cycle. However, our three conditions enforce the rule that no blocking edge leaves $V \setminus V'$. Therefore, the blocking edge must be in G_{P_Z}. These blocking edges can only be reached from dangerous vertices on an alternating path in G_P, and due to condition (b) above, those paths are all disrupted by a $(-, -)$ edge between V' and $V \setminus V'$.

2. G_P admits no alternating path that contains at least two blocking edges.
   As we have argued above, all blocking edges in G_P must also block P_Z. However, an entire alternating path with two blocking edges cannot run in V', because it would contradict the popularity of P_Z. For an alternating

path containing a blocking edge, leaving V' is also not an option, because of condition (b) above.

3. G_P admits no alternating path that contains a blocking edge and starts at a vertex in U.
   Similarly as above, the blocking edge must also block P_Z, and the alternating path would have to leave V', which is impossible due to condition (b) above.

With this we have shown that none of the three conditions in Theorem 1 are fulfilled by $P = P_Z \cup S$. □

4.3. Implementation of the second test

We now sketch the main idea on how to enforce the three conditions on edge labeling in Claim 3. For a pseudocode, please consult Algorithm 2.

The three conditions in Claim 3 list 3×2 edge labels in total: in condition (a), there is a $(+,-)$ label, in condition (b), there is a $(-,-)$ label, and finally, in condition (c), there is a $(+,-)$ label. The two $'+'$ signs among these edge labels are fulfilled, since the corresponding vertices are uncovered in P. Three of the four $'-'$ signs impose a requirement on vertices in $V \setminus V'$, while the fourth one imposes a requirement on vertices in D. This latter one can be found in condition (b), which requires $'-'$ votes for vertices in D on edges between D and $V \setminus V'$. Hence if a vertex $z \in D$ votes $'+'$ for a neighbor $x \in V \setminus V'$ then the stable matching S described in Claim 3 cannot exist. We test this property first, in lines 1 and 2, as it does not depend on how P_Z is completed to P. This constitutes the first phase of our subroutine.

For the completion of P_Z, we turn our attention to the remaining three $'-'$ signs in three edge deletion rounds of the second phase as follows. For a vertex x in $V \setminus V'$, each condition determines a list of edges incident to x such that the vote of x should be a $'-'$ for that edge with respect to P. Each of these requirements imposes an upper bound on the rank of possible partners for each vertex in $V \setminus V'$. Our key step , in lines 3 to 11 is to delete all the edges that are ranked worse than any of these bounds. Finally, in line 12 we test whether there is a stable matching in the remaining edge set covering every vertex in $V \setminus V'$. If so, then this is a suitable matching S, because none of the deleted edges would have blocked it, as every deleted edge has at least one vertex where the bound, and thus the matching partner is higher in the preferences than that edge.

Algorithm 2: Subroutine checking if a complete stable matching S exists on $V \setminus V'$ fulfilling the conditions in Claim 3

Input: A popular roommates problem instance and an initial matching P_Z that has passed the first test.

Output: A complete stable matching S on $V \setminus V'$ or a 'NO' answer.

for $z \in D$ *and* $x \in N(z) \cap (V \setminus V')$ **do**
  if $x \succ_z P_Z(z)$ **then**
    return *'NO'*

for $u \in U$ **do**
  for $x \in N(u) \cap (V \setminus V')$ **do**
    delete all (x, y) edges in $G[V \setminus V']$ such that $u \succ_x y$;

for $z \in D$ **do**
  for $x \in N(z) \cap (V \setminus V')$ **do**
    delete all (x, y) edges in $G[V \setminus V']$ such that $z \succ_x y$;

for $v \in V' \setminus (D \cup U)$ **do**
  for $x \in N(v) \cap (V \setminus V')$ *such that* $x \succ_v P_Z(v)$ **do**
    delete all (x, y) edges in $G[V \setminus V']$ such that $v \succ_x y$;

if *a complete stable matching S exists in the remaining graph* $G[V \setminus V']$ **then**
  return S

return *'NO'*

The correctness of Algorithm 2 is proven by the following claim.

Claim 4. *Assume that we are given a matching P_Z that has passed the first test in Algorithm 1 and the first phase of the second test. A matching S in the graph $G[V \setminus V']$ is stable, covers all vertices in $V \setminus V'$, and induces the three-condition edge-labeling criteria phrased in Claim 3 if and only if it is a stable matching that covers all vertices in $V \setminus V'$ in the graph we get after deleting each edge $(x, y) \in (V \setminus V') \times (V \setminus V')$ that fulfills any of the following conditions.*

1. $\exists u \in U$: $u \succ_x y$
2. $\exists z \in D$: $z \succ_x y$
3. $\exists v \in V' \setminus (D \cup U)$: $x \succ_v P_Z(v)$ *and* $v \succ_x y$

Proof. The condition on the set of covered vertices in the transformed graph is clearly necessary and sufficient, and it is straightforward that the matching in the new instance must be stable to ensure stability in the instance before edge deletions. We now revisit the three conditions in Claim 3 one-by-one and translate them into equivalent edge deletions. Claim 3 states that for each vertex $x \in V \setminus V'$, matching P must fulfill the following criteria:

1. (u, x) is a $(+, -)$ edge for $\forall u \in U$,
2. (z, x) is a $(-, -)$ edge for $\forall z \in D$,
3. (v, x) is a $(+, -)$ edge for $\forall v \in V' \setminus (D \cup U)$ such that $x \succ_v P(v)$.

Notice that the $'+'$ signs are unavoidable, while the first phase of Algorithm 2 guarantees the first $'-'$ sign in condition 2. Thus we only need to make sure that the conditions imposed by the remaining three $'-'$ signs are fulfilled. These state that each $x \in V \setminus V'$ is matched to a partner who is ranked better than each of 1) $u \in U$, 2) $z \in D$, and 3) $v \in V' \setminus (D \cup U)$ such that $x \succ_v P_Z(v)$. These conditions enforce the rule that all edges that are ranked worse by x than any of these three bounds need to be deleted to guarantee the three $'-'$ signs. In other words, an edge $(x, y) \in (V \setminus V') \times (V \setminus V')$ needs to be deleted if it fulfills any of the three conditions listed in Claim 4.

It is easy to see that if these three types of edge deletions are executed, then the edge labeling from Claim 3 is guaranteed. All that remains is to show is that stable matchings in the new instance are also stable in the original one. If such a matching M is blocked by an edge (x, y), then (x, y) must have been removed in the edge deletion round. However, due to the cover constraint on $V \setminus V'$, M must match at least one of x, y along an edge that is ranked better than (x, y). □

4.4. Correctness and running time

We first finish the proof of correctness for Algorithm 1, then estimate its running time. Finally, we prove the correctness and establish a running time estimate for utilizing Algorithm 1 for deciding whether a popular matching exists at all in an instance with high minimum degree and odd n.

Theorem 2. *For an independent vertex set U and $Z = V \setminus N(U) \setminus U$, a popular but unstable matching that leaves exactly the vertices in U uncovered or a proof for its non-existence is outputted by Algorithm 1 in $O\left(n^{|Z|} \cdot (m + |Z|^3)\right)$ time.*

Proof. We first prove that if Algorithm 1 terminates with a matching, then it is popular and leaves exactly the vertices in U uncovered. Let $P = P_Z \cup S$ be the outputted matching. By definition, P_Z covers the vertices in $Z \cup P_Z(Z)$, while S covers the vertices in $V \setminus V' = N(U) \setminus P_Z(Z)$, which is altogether $Z \cup N(U) = V \setminus U$. The popularity follows from Claim 3.

For the other direction we assume that P is a popular, but unstable matching in G, and that it leaves exactly U uncovered. Algorithm 1 tests all P_Z matchings that cover Z and are inclusion-wise minimal with respect to this property. We know from Corollary 1 that P is blocked by an edge that connects two vertices in $Z \cup P_Z(Z)$ for such a P_Z matching. This P_Z must therefore be generated in line 2, and pass the test in line 4. Furthermore, Claim 3 shows that $S = P \setminus P_Z$ exists, which then guarantees that line 5 is also passed for this particular P_Z. Note that more than one suitable stable matching might exist—our algorithm only outputs one stable matching that extends P_Z to a popular matching, which might be different from P itself.

Running time. We assume that the input is given in the form of a vertex set marking U and the strictly ordered list of edges at each vertex, in which list we assume comparison can be done in constant time. The lists also provide an edge list for the whole graph. The running time of Algorithm 1 is determined by the following factors.

1. Line 1: determine Z.

This takes $O(m)$, because it can be done by reading the list of edges once and deleting a vertex from V if it is in U, or adjacent to any vertex in U.

2. Line 2: determine $\mathcal{P}_Z$ to Z.
   Choosing a partner for each vertex in Z can be done in $O\left(\binom{n}{|Z|}\right)$ ways. Each of the feasible P_Z matchings constructed in this step needs to be investigated separately (line 3).

3. Line 4: check the popularity of P_Z in G'.
   We need to construct the edge labeling with respect to P_Z in $O(m)$ time and then check whether P_Z is indeed a maximum weight matching in the instance. This can be done in $O(|Z|^3)$ time (Schrijver, 2003, Theorem 26.2).

4. Line 4: check the instability of P_Z in G'.
   The labeling in the previous step already locates all blocking edges.

5. Line 5: checking the conditions in Claim 3 by calling Algorithm 2.

   - Determining D can be done by growing alternating trees starting from each blocking edge in G'_{P_Z}. There are at most $O(|Z|)$ end vertices of blocking edges, and even checking all paths starting from them can be done in $O(|Z|^3)$ time (Schrijver, 2003, Section 24.2a).
   - Testing in the first phase of Algorithm 2 can be done in constant time. Also, based on our three conditions in Claim 4, for each edge it can be decided in constant time whether it should be deleted or not.
   - Irving's algorithm (Irving, 1985) can find a stable matching or a proof for its non-existence in the transformed graph in $O(m)$ time. If the found stable matching does not cover all vertices in $V \setminus V'$, then no stable in this instance does, due to the Rural Hospitals Theorem (Gusfield & Irving, 1989, Theorem 4.5.2).

In total, this yields the following running time.

$$O\left(m + \binom{n}{|Z|} \cdot (m + |Z|^3 + |Z|^3 + m + m)\right) \longrightarrow O\left(n^{|Z|} \cdot (m + |Z|^3)\right) \quad \square$$

Algorithm 1 can be utilized on a larger scale, as the following theorem shows.

Theorem 3. *If graph G has an odd n and $\deg(v) \geq n - c$ for all $v \in V$ and some constant c, the repeated running of Algorithm 1 can output a popular matching or a proof for its non-existence in $O\left(n^c \cdot (m + c^3)\right)$ time.*

Proof. Algorithm 1 can test on any instance of the popular roommates problem whether there is a popular matching that leaves a given non-empty independent vertex set U uncovered. In graphs on an odd number of vertices, all matchings leave a non-empty set of vertices uncovered. Therefore, if n is odd, we can iterate through all possible U sets and derive whether a popular matching exists in the instance at all. For an odd n and $\deg(v) \geq n - c$ for all $v \in V$ and some constant c, from the definition of Z follows that $|Z| < c$. Since no vertex $u \in U$ is adjacent to a vertex in $U \cup Z$, we have that $|U \cup Z| \leq c$. Thus, the running time for checking the existence of a popular matching is polynomial for a constant c, as the following calculation demonstrates.

$$O\left(\sum_{i=1}^{c} \binom{n}{i} \cdot n^{c-i} \cdot (m + (c-i)^3)\right) \rightarrow O\left(n^c \cdot (m + c^3)\right) \qquad \square$$

We also remark that by iterating through all possible sets of uncovered vertices in an increasing order of $|U|$, our algorithm is able to find a maximum size popular matching as well in graphs with an odd n and $\deg(v) \geq n - c$ for all $v \in V$. For graphs with an even n, our algorithm is only able to decide whether a non-perfect popular matching exists. This is not surprising, because deciding whether a (perfect) popular matching exists in a complete instance with an even n is NP-complete (Cseh & Kavitha, 2021), while our algorithm runs in polynomial time if $|U|$ is small and the minimum degree in the graph is large.

5. EXAMPLE

We demonstrate our algorithm on a chosen P_Z in the instance depicted in Figure 3. Here we only discuss the case $U = \{b\}$, $P_Z = \{(e, f), (g, h)\}$. A fully detailed example that checks the existence of a popular matching in the same instance can be found in the appendix.

For this P_Z, the graph $G' = G[Z \cup P_Z(Z) \cup U]$ is defined by vertices $V' = \{b, e, f, g, h\}$ and edges $\{(e, f), (e, g), (f, g), (g, h)\}$. The first test in Algorithm 1 checks whether P_Z is popular in G'. As only one edge in G', namely (e, g), blocks P_Z and from this edge there is no alternating path or

$a:\ d,\ b,\ e$
$b:\ a,\ d$
$d:\ b,\ a,\ e,\ h$
$e:\ g,\ a,\ d,\ f$
$f:\ g,\ e$
$g:\ e,\ h,\ f$
$h:\ g,\ d$

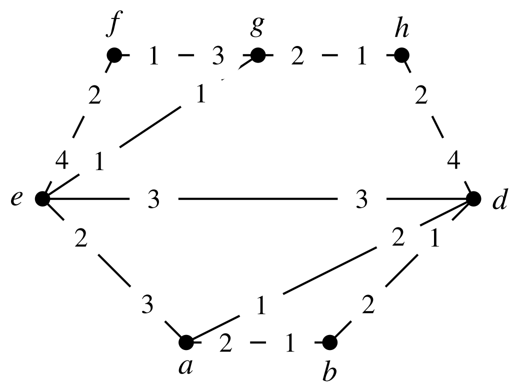

Figure 3: An example instance with no stable matching and no popular matching.

cycle in G'_{P_Z}, based on the characterization in Theorem 1 we can conclude that P_Z is popular in G'. We remark that in practice, the fulfillment of the characterization is checked such that the three conditions are converted into the weight requirement in a weighted matching instance defined specifically for P_Z—see Section 2. The first test in Algorithm 1 furthermore requires the existence of a blocking edge for P_Z in G', which is granted by (e, g). The first test is thus passed by P_Z.

Consider now the second test in Algorithm 1. For this test we build the set of remaining vertices $V \setminus V' = \{a, d\}$ and the set of dangerous vertices $D = \{f, h\}$. Vertex f has no neighbor in $V \setminus V'$, h is adjacent to d, and $P_Z(h) \succ_h d$, so the stable matching (a, d) passes the first phase of Algorithm 2 in lines 1-2. However, in line 5 of Algorithm 2 this edge is deleted, because $d \in N(b) \cap (V \setminus V')$ and $b \succ_d a$. Therefore the remaining graph on $V \setminus V'$ is the empty graph on the two vertices a and d, and there is no complete matching that covers these two vertices after the edge deletion steps, so Algorithm 2 returns the answer ‘NO’.

We conclude that this P_Z fails the second test in Algorithm 1 because on $V \setminus V'$ there is no complete stable matching after the edge deletions.

6. EXPERIMENTAL STUDY

We tested our algorithm on various randomly generated instances. For each combination of $n \in \{7, 9, 11, 13, 15\}$ and $c \in \{3, 4, 5, 6\}$, 100 000 graphs of

	$n = 7$	$n = 9$	$n = 11$	$n = 13$	$n = 15$
$c = 3$	60 890 18	47 757 3	38 578 0	31 490 0	26 274 0
$c = 4$	72 848 175	56 160 21	44 763 5	36 122 1	29 823 0
$c = 5$	85 585 961	65 961 133	51 952 21	41 547 6	33 884 1
$c = 6$	96 234 1 641	76 838 593	60 203 83	56 292 15	38 857 4

Table 1: The tested n and c values are organized in the different columns and rows. Each cell contains the number of instances that admit a popular matching as the top entry and the number of instances that admit a popular, but no stable matching as the bottom entry. The number of instances that admit a stable matching is the difference of the two numbers in each cell. These numbers are consistent with results measured in earlier studies on the number of instances admitting a stable matching (Irving, 1985; Prosser, 2014).

minimum degree $n - c$ were generated using the Erdős–Rényi model (Erdős & Rényi, 1960). Each edge was added with probability $p = \frac{n-c+1}{n-1}$, as this guarantees that among the $n - 1$ possible edges of a vertex, more than $n - c$ are chosen in expectation. Only those graphs were kept that met the bound $n - c$ on the minimum degree and indeed had at least one vertex of degree exactly $n - c$. The preference list of each vertex was then generated by choosing uniformly at random among all permutations of its edges.

We first ran Irving's algorithm (Irving, 1985) on each generated instance to test whether it admits a stable matching, and for those with no stable matching, we also ran our algorithm for all odd-cardinality independent vertex sets U with $|U| \leq c - 2$ to test whether there is a popular matching leaving exactly U uncovered. These tests delivered the number of instances with a stable (and thus popular) matching, and the number of instances with a popular, but no stable matching among the 100 000 tested instances. The entries in Table 1 are calculated from these numbers for all tested (n, c) combinations.

Focusing on one row of the table, one can observe that for a fixed c and increasing n, the number of instances that admit a popular matching decreases. A reason for this might be that the probability of an improving path or cycle (see Theorem 1) increases as the graph grows. Fixing n and increasing c

leads to more instances that admit a popular matching. Analogously to the previous reasoning, this might happen because allowing more sparse graphs gives less room for improving paths or cycles. In general, we can observe that for randomly generated graphs with a high minimum degree, very few instances occur that admit a popular matching but no stable matching. As the minimum degree requirement is gradually relaxed, their number, and also their ratio among instances that admit a popular matching at all, increases. We remind the reader that for an odd n and $c \leq 2$, a matching is popular if and only if it is stable (Cseh & Kavitha, 2021).

The experiments were run on a standard desktop computer powered by Intel i9-11900F CPU running at 5.2GHz and 16 GiB of RAM. We used the LEMON Graph Library (Dezső et al., 2011) implementation of the maximum weight perfect matching algorithm, which is based on the blossom algorithm of Edmonds (1965). Our source code is available on the website.[1]

For the $n = 7, c = 3$ setting, deciding the existence of a popular matching took 4 seconds for those instances that admit no stable matching out of the 100 000 instances in total. For comparison, we generated all matchings on the same set of instances and checked each one of them for popularity, which took 37 seconds. For $n = 11, c = 3$, these runtimes increased to 139 seconds and close to 650 hours, respectively. These numbers demonstrate that for graphs with a high minimum degree, our algorithm is indeed much faster than brute-forcing through all matchings to check whether the instance with no stable matching admits a popular matching.

7. RELATION TO OTHER DECOMPOSITIONS AND OPEN QUESTIONS

In this section we first compare our approach with two earlier decomposition techniques, and then list some open questions.

7.1. Other decompositions of popular matchings

Decomposing a popular matching to a set of edges that is stable in their subgraph and to a set of edges that is popular in their subgraph appears in two further papers.

[1] See https://github.com/atimaly/Stable-and-Popular-Matchings.

1. Cseh & Kavitha (2021) identified the set of edges that can appear in a popular matching in a bipartite instance by showing that any popular matching M can be decomposed as $M = M_0 \cup M_1$, where M_0 is a so-called *dominant* matching in the subgraph induced by the vertices matched in M_0, and in the subgraph induced by the remaining vertices, M_1 is stable. A popular matching is called dominant if it is more popular than any larger matching.

2. Kavitha (2024) searched for popular matchings in non-bipartite instances by showing that every popular matching can be partitioned into a stable part and a *truly popular* part. A matching M is defined as truly popular if it is a popular fractional matching.

Our decomposition is substantially different from both of these techniques, as argued below.

1. A complete popular matching is always dominant, thus our P_Z is a dominant matching within $V(P_Z)$. Our decomposition is similar to the one in Cseh & Kavitha (2021) in this view, however, the main property we use is that *all* extensions of a given matching P_Z to a popular matching are *exactly* the set of complete stable matchings of a given auxiliary stable matching instance, which does not necessarily hold in Cseh & Kavitha (2021).

2. Our popular part P_Z is not necessarily truly popular within $V(P_Z)$. We demonstrate this in Figure 4, using the example given in (Kavitha, 2024, Figure 2). For set $U = \{u\}$ we have $Z = \{a, b, c, e, f\}$. Matching $P_Z = P = \{(a, b), (c, d), (e, f)\}$, marked by dashed blue edges, is popular and blocked by the dotted red edges $\{(d, e), (d, f)\}$. Then matching P_Z in $V(P_Z)$ is exactly the example in (Kavitha, 2024, Figure 2), which is not truly popular.

7.2. Open questions

The most promising direction might be to accelerate our algorithm to reach a fixed parameter tractability result for the discussed case or to discover further polynomially solvable instance classes of the popular roommates problem. As argued in Section 4.4, graphs with a high minimum degree can only be

$a: \ b, \ c$
$b: \ c, \ a$
$c: \ d, \ a, \ b$
$d: \ e, \ f, \ c, \ u$
$e: \ d, \ f$
$f: \ d, \ e$
$u: \ d$

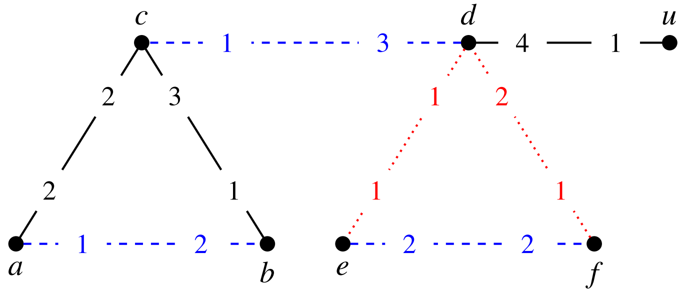

Figure 4: An example for a popular P_Z that is not truly popular.

a subject of such investigation if combined with an odd n, which is the case we covered in this paper. However, other graph parameters might prove to be fruitful. Parameterized complexity results on other popular matching problems restrict the variability of preferences (Huang et al., 2011; Krishnapriya et al., 2018) or operate on graphs with a bounded treewidth (Faenza et al., 2019).

A. EXAMPLES

We now list two examples, one NO-instance and one YES-instance.

A.1. NO-instance

We demonstrate a full run of our algorithm on the instance depicted in Figure 3. It is easy to see that the instance admits no stable matching, because each stable matching would have to include the edge (e, g), and after fixing this edge, we are left with a preference cycle of length 3 on vertices $\{a, b, d\}$, which proves the non-existence of a stable matching.

Since n is odd, each popular matching must leave an odd number of vertices uncovered. In order to check whether there is a popular matching, we now iterate through all vertex sets U with an odd cardinality and no spanned edge, and apply Algorithms 1 and 2 to determine whether there is a popular matching that leaves exactly the vertices in U uncovered. Table 2 summarizes these steps. As the instance admits no popular matching, each tested P_Z fails at some point. The exact explanation on how this happens is deferred to the list marked with capital letters.

Case	P_Z	$V(G')$	$E(G')$	test
$U = \{a\}$	$(f,g),(d,h)$	a,d,f,g,h	$(a,d),(d,h),(f,g),(g,h)$	1 (A)
$Z = \{f,g,h\}$	$(e,f),(g,h)$	a,e,f,g,h	$(a,e),(e,f),(e,g),(f,g),(g,h)$	1 (B)
$U = \{b\}$	$(e,f),(g,h)$	b,e,f,g,h	$(e,f),(e,g),(f,g),(g,h)$	2 (C)
$Z = \{e,f,g,h\}$	$(a,e),(f,g),(d,h)$	$V(G)$	$E(G)$	1 (D)
$U = \{d\}$	(f,g)	d,f,g	(f,g)	1 (E)
$Z = \{f,g\}$	$(e,f,(g,h)$	d,e,f,g,h	$(d,e),(d,h),(e,f),(e,g),(f,g),(g,h)$	1 (F)
$U = \{e\}$	$(a,b),(d,h)$	a,b,d,e,h	$(a,b),(a,d),(a,e),(b,d),(d,e),(d,h)$	1 (F)
$Z = \{b,h\}$	$(a,b),(g,h)$	a,b,e,g,h	$(a,b),(a,e),(e,g),(g,h)$	1 (G)
	$(b,d),(g,h)$	a,b,e,g,h	$(a,b),(a,e),(e,g),(g,h)$	1 (G)
$U = \{f\}$	$(a,b),(d,h)$	a,b,d,f,h	$(a,b),(a,d),(b,d),(d,h)$	2 (H)
$Z = \{a,b,d,h\}$	$(a,b),(d,e),(g,h)$	$V(G)$	$E(G)$	1 (I)
	$(a,e),(b,d),(g,h)$	$V(G)$	$E(G)$	1 (J)
$U = \{g\}$	$(a,b),(d,e)$	a,b,d,e,g	$(a,b),(a,d),(a,e),(b,d),(d,e),(e,g)$	1 (G)
$Z = \{a,b,d\}$	$(a,b),(d,h)$	a,b,d,g,h	$(a,b),(a,d),(b,d),(d,h),(g,h)$	1 (G)
	$(a,e),(b,d)$	a,b,d,e,g	$(a,b),(a,d),(a,e),(b,d),(d,e),(e,g)$	1 (G)
$U = \{h\}$	$(a,b),(e,f)$	a,b,e,f,h	$(a,b),(a,e),(e,f)$	1 (E)
$Z = \{a,b,e,f\}$	$(a,b),(d,e),(f,g)$	$V(G)$	$E(G)$	1 (K)
	$(a,e),(b,d),(f,g)$	$V(G)$	$E(G)$	1 (L)
$U = \{a,f,h\}$	$\emptyset$	a,f,h	$\emptyset$	2 (M)
$Z = \emptyset$				
$U = \{b,e,h\}$	$\emptyset$	b,e,h	$\emptyset$	2 (N)
$Z = \emptyset$				
$U = \{b,f,h\}$	$\emptyset$	b,f,h	$\emptyset$	2 (O)
$Z = \emptyset$				

Table 2: The first column contains the tested U set and the set of vertices Z, calculated from U. In the second column, we list the possible P_Z matchings, one per row. To each of these, the vertices and edges of G' are listed. Finally, we mark which test the current P_Z failed and give a detailed explanation of this failure in a separate list.

A: P_Z is not popular in G', because the blocking edge (g,h) can be reached from a in G'_{P_Z}.

B: P_Z is not popular in G', as (a,e) blocks P_Z.

C: Edge (e,g) blocks P_Z and $D = \{f,h\}$. The graph spanned by $V \setminus V' = \{a,d\}$ consists of the edge (a,d) itself, but even this edge is deleted in Algorithm 2. We conclude that after the edge deletions in the second test, no stable matching covers the entire $V \setminus V'$.

D: P_Z is not popular in G', as (b,d) blocks P_Z.

E: P_Z is stable in G'.

F: P_Z is not popular in G', as (d, e) blocks P_Z.

G: P_Z is not popular in G', as (e, g) blocks P_Z.

H: Edge (a, d) blocks P_Z and $D = \{b, h\}$. The graph spanned by $V \setminus V' = \{e, g\}$ consists of the edge (a, d) itself, but even this edge is deleted in Algorithm 2, because $g \succ_h d$. We conclude that after the edge deletions in the second test, no stable matching covers the entire $V \setminus V'$.

I: P_Z is not popular in G', because the blocking edge (a, d) can be reached from f in G'_{P_Z}.

J: P_Z is not popular in G', because the blocking edge (a, b) can be reached from f in G'_{P_Z}.

K: P_Z is not popular in G', because the blocking edge (e, g) can be reached from h in G'_{P_Z}.

L: P_Z is not popular in G', because the blocking edge (a, b) can be reached from h in G'_{P_Z}.

M: In Algorithm 2 we delete (b, d) since $a \succ_b d$. After this deletion, no matching covers $V \setminus V' = \{b, d, e, g\}$.

N: In Algorithm 2 we delete (a, d) since $b \succ_d a$. After this deletion, no matching covers $V \setminus V' = \{a, d, f, g\}$.

O: In Algorithm 2 we delete (a, d) since $b \succ_d a$. After this deletion, no matching covers $V \setminus V' = \{a, d, e, g\}$.

A.2. YES-instance

The instance depicted in Figure 5 admits a popular matching, but no stable matching. Since n is odd, each popular matching must leave an odd number of vertices uncovered. In order to check whether there is a popular matching, we start iterating through all vertex sets U with an odd cardinality and no spanned edge, and apply Algorithms 1 and 2 to determine whether there is a popular matching that leaves exactly the vertices in U uncovered. Table 3 summarizes these steps. A popular matching consisting of edges $(a, b), (d, h), (e, g)$ is found in the line marked by (*). The algorithm then terminates with outputting this matching.

$$\begin{array}{llllll}
a: & e, & b & & & \\
b: & e, & h, & g, & d, & a, & f \\
d: & h, & e, & b & & \\
e: & d, & h, & b, & g, & a, & f \\
f: & b, & e & & & \\
g: & e, & b, & h & & \\
h: & g, & e, & d, & b &
\end{array}$$

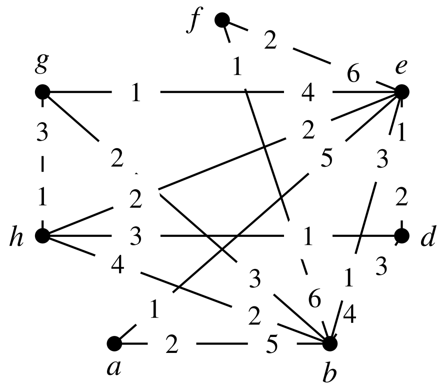

Figure 5: Example instance on 7 vertices that admits a popular matching but no stable matching.

Case	P_Z	$V(G')$	$E(G')$	test	reason for test failure (if any)
$U = \{a\}$	$(b,d),(e,f),(g,h)$	$V(G)$	$E(G)$	1	(a,e) blocks P_Z
$Z = \{d,f,g,h\}$	$(b,f),(d,e),(g,h)$	$V(G)$	$E(G)$	1	(a,e) blocks P_Z
	$(b,f),(d,h),(e,g)$	$V(G)$	$E(G)$	1	(a,b) blocks P_Z
	$(b,g),(d,h),(e,f)$	$V(G)$	$E(G)$	1	(a,b) blocks P_Z
$U = \{b\}$ $Z = \emptyset$	$\emptyset$	b	$\emptyset$	2	no complete matching in $G \setminus G'$
$U = \{d\}$	$(a,b),(e,f),(g,h)$	$V(G)$	$E(G)$	1	(d,e) blocks P_Z
$Z = \{a,f,g\}$	$(a,e),(b,f),(g,h)$	$V(G)$	$E(G)$	1	(d,e) blocks P_Z
$U = \{e\}$ $Z = \emptyset$	$\emptyset$	e	$\emptyset$	2	no complete matching in $G \setminus G'$
$U = \{f\}$ $Z = \{a,d,g,h\}$	$(a,b),(d,e),(g,h)$	$V(G)$	$E(G)$	1	the blocking edge (b,g) is reachable from f in G_{P_Z} via an alternating path
	$(a,e),(b,d),(g,h)$	$V(G)$	$E(G)$	1	the blocking edge (d,e) is reachable from f in G_{P_Z} via an alternating path
	$(a,e),(b,g),(d,h)$	$V(G)$	$E(G)$	1	the blocking edge (e,g) is reachable from f in G_{P_Z} via an alternating path
	$(a,b),(d,h),(e,g)$	$V(G)$	$E(G)$	-	*

Table 3: The first column contains the tested U set and the set of vertices Z, calculated from U. In the second column, we list the possible P_Z matchings, one per row. To each of these, the vertices and edges of G' are listed. Then we mark which test the current P_Z failed and explain this failure in the last column.

References

Abraham, D. J., Irving, R. W., Kavitha, T., & Mehlhorn, K. (2007). Popular matchings. *SIAM Journal on Computing*, *37*, 1030-1045.

Biró, P., Irving, R. W., & Manlove, D. F. (2010). Popular matchings in the marriage and roommates problems. In *Proceedings of CIAC '10: The 7th International Conference on Algorithms and Complexity* (Vol. 6078, p. 97-108). Springer.

Brandl, F., & Kavitha, T. (2019). Two problems in max-size popular matchings. *Algorithmica*, *81*, 2738–2764.

Chung, K. S. (2000). On the existence of stable roommate matchings. *Games and Economic Behavior*, *33*(2), 206-230.

Condorcet, M. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'Imprimerie Royale.

Cseh, Á. (2017). Popular matchings. *Trends in Computational Social Choice*, *105*.

Cseh, Á., Faenza, Y., Kavitha, T., & Powers, V. (2022). Understanding popular matchings via stable matchings. *SIAM Journal on Discrete Mathematics*, *36*(1), 188–213.

Cseh, Á., & Kavitha, T. (2021). Popular matchings in complete graphs. *Algorithmica*, *83*(5), 1493–1523.

Dezső, B., Jüttner, A., & Kovács, P. (2011). LEMON – an open source C++ graph template library. *Electronic Notes in Theoretical Computer Science*, *264*(5), 23-45. (Proceedings of the Second Workshop on Generative Technologies (WGT) 2010)

Edmonds, J. (1965). Paths, trees, and flowers. *Canadian Journal of Mathematics*, *17*, 449–467.

Erdem, E., Fidan, M., Manlove, D., & Prosser, P. (2020). A general framework for stable roommates problems using answer set programming. *Theory and Practice of Logic Programming*, *20*(6), 911–925.

Erdős, P., & Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, *5*(1), 17–60.

Faenza, Y., Kavitha, T., Powers, V., & Zhang, X. (2019). Popular matchings and limits to tractability. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 2790–2809).

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

Gärdenfors, P. (1975). Match making: Assignments based on bilateral preferences. *Behavioural Science*, *20*, 166-173.

Gupta, S., Misra, P., Saurabh, S., & Zehavi, M. (2021). Popular matching in roommates setting is NP-hard. *ACM Transactions on Computation Theory (TOCT)*, *13*(2), 1–20.

Gusfield, D., & Irving, R. W. (1989). *The Stable Marriage Problem: Structure and*

Algorithms. MIT Press.

Hirakawa, M., Yamauchi, Y., Kijima, S., & Yamashita, M. (2015). On the structure of popular matchings in the stable marriage problem — who can join a popular matching. In *The 3rd International Workshop on Matching under Preferences (MATCH-UP).*

Huang, C.-C., & Kavitha, T. (2013a). Near-popular matchings in the roommates problem. *SIAM Journal on Discrete Mathematics*, *27*(1), 43-62.

Huang, C.-C., & Kavitha, T. (2013b). Popular matchings in the stable marriage problem. *Information and Computation*, *222*, 180-194.

Huang, C.-C., & Kavitha, T. (2021). Popularity, mixed matchings, and self-duality. *Mathematics of Operations Research*, *46*(2), 405-427.

Huang, C.-C., Kavitha, T., Michail, D., & Nasre, M. (2011). Bounded unpopularity matchings. *Algorithmica*, *61*(3), 738-757.

Irving, R. W. (1985). An efficient algorithm for the "stable roommates" problem. *Journal of Algorithms*, *6*, 577-595.

Itoh, T., & Watanabe, O. (2010). Weighted random popular matchings. *Random Structures and Algorithms*, *37*(4), 477-494.

Kavitha, T. (2014). A size-popularity tradeoff in the stable marriage problem. *SIAM Journal on Computing*, *43*, 52-71.

Kavitha, T. (2024). Popular roommates in simply exponential time. *Algorithmica*, 1–29.

Krishnapriya, A. M., Nasre, M., Nimbhorkar, P., & Rawat, A. (2018). How good are popular matchings? In *17th International Symposium on Experimental Algorithms (SEA 2018)* (Vol. 103, p. 9).

Mahdian, M. (2006). Random popular matchings. In J. Feigenbaum, J. C. Chuang, & D. M. Pennock (Eds.), *Proceedings of EC '06: the 7th ACM Conference on Electronic Commerce* (p. 238-242). ACM.

Manlove, D. F. (2013). *Algorithmics of Matching under Preferences*. World Scientific.

Mertens, S. (2015). Stable roommates problem with random preferences. *Journal of Statistical Mechanics: Theory and Experiment*, *2015*(1), P01020.

Palmer, C., & Pálvölgyi, D. (2022). At most 3.55^n stable matchings. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)* (pp. 217–227).

Pittel, B. G. (1993). The stable roommates problem with random preferences. *Annals of Probability*, *21*(3), 1441-1477.

Pittel, B. G., & Irving, R. W. (1994). An upper bound for the solvability probability of a random stable roommates instance. *Random Structures and Algorithms*, *5*, 465-486.

Prosser, P. (2014). Stable roommates and constraint programming. In *International Conference on AI and OR Techniques in Constriant Programming for Combinatorial*

Optimization Problems (pp. 15–28).

Ruangwises, S., & Itoh, T. (2019). Random popular matchings with incomplete preference lists. *Journal of Graph Algorithms and Applications*, *23*(5), 815–835.

Schrijver, A. (2003). *Combinatorial Optimization: Polyhedra and Efficiency*. Springer.

Journal of Mechanism and Institution Design
ISSN: 2399-844X (Print), 2399-8458 (Online)
DOI:10.22574/jmid.2025.12.004

ON THE INFLUENCE OF NETWORK STRUCTURE ON THE RESILIENCE AND LOSSES OF FINANCIAL SYSTEMS

Lorenzo Zino, Politecnico di Torino, Italy
lorenzo.zino@polito.it

Giulia Fracastoro, Politecnico di Torino, Italy
giulia.fracastoro@polito.it

Anton V. Proskurnikov, Politecnico di Torino, Italy
anton.p.1982@ieee.org

Giuseppe C. Calafiore, Politecnico di Torino, Italy
giuseppe.calafiore@polito.it

ABSTRACT

We investigate resilience to defaults in financial networks, where nodes hold shares in a common set of external assets. Price fluctuations in these assets can trigger shareholder defaults, propagating through the network and causing default cascades. In particular, we conduct a series of numerical experiments to elucidate the influence of the network structure on the financial system. Our investigation explores the effects of sparsity versus connectivity, clustering, and liability heterogeneity among financial institutions. We also examine the impact of diversification and asset investment heterogeneity.

Keywords: Default cascades, financial networks, resilience.

JEL Classification Numbers: C02, C61, G01.

This study was carried out within the 2022K8EZBW "Higher-order interactions in social dynamics with application to monetary networks" project — funded by European Union — Next Generation EU within the PRIN 2022 program (D.D. 104 — 02/02/2022 Ministero dell'Università e della Ricerca). This manuscript reflects only the authors' views and opinions and the Ministry cannot be considered responsible for them.

1. INTRODUCTION

In recent decades, scientific interest in the vulnerability of financial networks and related systemic risk factors has grown, particularly in the aftermath of the 2007–2008 financial crisis. Understanding complex liability networks among financial institutions and their impact on the resilience of a financial system to external shocks is crucial for preventing default cascades. Despite numerous efforts, this impact remains insufficiently explored. On the one hand, a dense liability network can enhance resilience to external shocks by providing more connections to absorb a moderate shock (Allen & Gale, 2000; Freixas et al., 2000; Babus, 2016). On the other hand, dense financial networks are more fragile to intense shocks once a critical number of nodes have been affected, as more connections create additional paths for default contagion (Glasserman & Young, 2016; Acemoglu et al., 2015; Nier et al., 2007; Haldane & May, 2011; Bardoscia et al., 2017; Elliott et al., 2014).

Various models have been developed to investigate default propagation in financial networks, many inspired by the seminal work in Eisenberg & Noe (2001), which laid the foundation for extensive studies on financial contagion (Glasserman & Young, 2016; Elsinger et al., 2006; Rogers & Veraart, 2013; Acemoglu et al., 2015; Kusnetsov & Veraart, 2019; Feinstein, 2017; Hurd, 2016; Massai et al., 2022). Their model addresses the clearing of mutual liabilities when institutions default due to external shocks, providing an explicit method to compute the clearing payment matrix and total shortfall. Extensions of the Eisenberg-Noe model have incorporated crucial features of financial systems, such as prioritizing liabilities to the non-financial sector over interbank liabilities (Elsinger et al., 2006).

A critical aspect of systemic risk is asset commonality within financial networks. Alongside mutual liabilities, financial institutions hold shares in external assets. Price fluctuations in these assets can trigger complex liquidity shocks, affecting all shareholders simultaneously (Cifuentes et al., 2005; Glasserman & Young, 2016; Haldane & May, 2011; Allen et al., 2012; Banerjee & Feinstein, 2022; Amini & Feinstein, 2023), and potentially amplifying through interbank connections. Initial defaults of banks exposed to depreciating assets may trigger secondary defaults, impacting even those without direct exposure. This occurs when banks experience equity reductions due to unpaid liabilities from defaulted banks, hindering their ability to meet obligations. Recent work in Calafiore et al. (2025) quantifies the worst-case impact of such default chains, which can lead to significant losses across the financial system. Using linear programming duality, Calafiore et al. (2025) identifies the maximum price fluctuation level that ensures no bank defaults and all obligations are fulfilled. For price fluctuations exceeding this threshold, the maximum possible loss to the financial system is evaluated. Additionally, Calafiore et al. (2025) determines a critical threshold for asset price variations below which no bank becomes insolvent.

In this paper, building on the theoretical tools developed in Calafiore et al. (2025), we investigate how the structural properties of financial networks influence their resilience to shocks, focusing on the impact of network structures on worst-case system loss during price fluctuations. By structural properties, we refer to the topologies of the interbank liability graph, which depicts mutual liabilities among financial institutions, and the institution-to-asset graph, which captures how institutions diversify their investments across assets. To address this, we perform Monte Carlo-based numerical experiments to incorporate the inherent stochasticity of financial systems. These experiments allow us to analyze the effects of network structures on resilience margins and worst-case system loss, as defined in Calafiore et al. (2025), under various scenarios. Our main contributions are as follows.

First, we investigate the role of interbank connection density by comparing performance across different interbank networks, modeled as random regular digraphs with increasing degree (where each institution has liabilities to a fixed number of others). Our results indicate that very sparse interbank networks are more susceptible to default cascades; however, the benefits of increased connectivity quickly reach a saturation point. This aligns with existing literature, which highlights two opposing effects of increasing network connectivity that appear to balance each other in this scenario (Glasserman & Young, 2016; Acemoglu et al., 2015; Bardoscia et al., 2017). Second, we examine the role of the clustering coefficient by comparing performance on Watts–Strogatz networks with varying re-wiring probabilities, and thus different levels of clustering (Newman, 2018), concluding that the clustering coefficient has a negligible effect on resilience.

Third, we examine the role of heterogeneity in the interbank network by comparing results on random regular graphs with those on irregular topologies, where institutions have liabilities to varying numbers of others. Interestingly, our findings reveal a nontrivial and nonmonotone behavior: lower levels of heterogeneity appear to enhance the resilience of the financial network, while excessive heterogeneity has the opposite effect. These observations align with findings in some other model-based analyses (Caccioli et al., 2012; Battiston et al., 2012; Iori et al., 2006). Finally, we examine the network of institutional investments across different assets, analyzing how its connectivity (representing diversification) and heterogeneity impact the resilience of the financial system. As expected, our numerical results show that diversification consistently benefits financial stability. However, varying levels of diversification appear to weaken the system, indicating that a minority of institutions with poorly diversified assets can jeopardize the overall functioning of the financial network.

The rest of the paper is organized as follows. In Section 2, we present the mathematical model used for financial networks and present our research problem. Section 3 describes our main results. Section 4 concludes the paper and outlines possible future research. We summarize the key mathematical results used in resilience analysis in APPENDIX A.

2. FINANCIAL NETWORKS AND PROBLEM SETUP

We denote the set of strictly positive integers, nonnegative real numbers, and strictly positive real numbers as $\mathbb{N}_{>0}$, $\mathbb{N}_{\geq 0}$, and $\mathbb{R}_{>0}$, respectively. A vector $\boldsymbol{x}$ is represented in bold lowercase, with x_i as its ith entry. A matrix A is represented in uppercase, with a_{ij} as the jth entry of its ith row. We denote the all-0 and the all-1 vectors as $\mathbf{0}$ and $\mathbf{1}$, respectively, and the identity matrix as I, where dimensions are omitted when unnecessary. For vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of the same dimension, $\boldsymbol{x} \geq \boldsymbol{y}$ denotes the component-wise inequalities, $x_i \geq y_i$ for all i. The relation $\leq$ is defined analogously. The operations max and min are defined element-wise; the positive part of a vector $\boldsymbol{x}$ is defined as $\boldsymbol{x}^+ = \max\{\boldsymbol{x}, \mathbf{0}\}$.

2.1. Interbank Claims Network

We consider a (finite) set $\mathcal{V}$ of $n \geq 2$ financial institutions (hereafter called *banks* for simplicity). Each bank $i \in \mathcal{V}$ has a net position with respect to the non-financial sector (referred to as its *external net position*), denoted by $c_i \in \mathbb{R}$, which represents all the bank liquidity, assets, and debts with the external sectors. The external net positions of all banks are represented by the vector $\boldsymbol{c} \in \mathbb{R}^n$. The net position can be positive or negative (Elsinger et al., 2006; Hurd, 2016), since it ultimately represents the difference between the positive external assets and debts. Note that this assumption differs from the original Eisenberg-Noe model (Eisenberg & Noe, 2001), where it is assumed that $c_i \geq 0$.

Besides external assets and liabilities, each bank may have obligations to other banks, captured by a $n \times n$ nonnegative-valued matrix $\bar{P} \in \mathbb{R}^{n\times n}_{\geq 0}$ of nominal liabilities (for instance, bank i is obligated to pay $\bar{p}_{ij}$ currency units to bank j). By definition, $\bar{p}_{ii} = 0$ (banks have no self-liability). The total liability $\bar{p}_i$ of bank i to other banks is obtained by summing the elements on the ith row of matrix $\bar{P}$, i.e.,

$$\bar{p}_i := \sum_{j\in\mathcal{V}} \bar{p}_{ij}. \tag{1}$$

Gathering these values into the vector $\bar{\boldsymbol{p}} \in \mathbb{R}^n_{\geq 0}$, one has $\bar{\boldsymbol{p}} = \bar{P}\mathbf{1}$. Similarly, the sum of the ith column of matrix $\bar{P}$, $\sum_{k\in\mathcal{V}} \bar{p}_{ki}$, represents the total liability of other banks to bank i. We introduce the stochastic matrix of *relative liabilities* $A \in [0,1]^{n\times n}$, where each entry is defined as

$$a_{ij} = \begin{cases} \frac{\bar{p}_{ij}}{\bar{p}_i} & \text{if } \bar{p}_i > 0, \\ 1 & \text{if } \bar{p}_i = 0 \text{ and } i = j, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

and represents the fraction of bank i's total liability $\bar{p}_i$ that is owed to bank j.

Such interconnections between banks naturally induce a weighted directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \bar{P})$, where a (directed) edge $(i, j) \in \mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is present if and only if i has an obligation to bank j, i.e., if and only if $\bar{p}_{ij} > 0$. See, e.g., the green connections in Fig. 1. We observe that, since $\bar{P}$ is a zero-diagonal matrix, the graph $\mathcal{G}$ has no self-loops. We shall refer to the graph $\mathcal{G}$ as an *interbank network*.

2.2. Bank-to-Asset Network

Following the approach in Calafiore et al. (2025), we assume that the external net position comprises two components. First, each bank $i \in \mathcal{V}$ may hold net liquidity, denoted by c^e_i, with these values collected in the vector $\boldsymbol{c}^e := [c^e_1, \ldots, c^e_n]^\top$. Second, each bank may own shares in external assets. Regarding the latter, we assume a set of $m \in \mathbb{N}_{>0}$ liquid assets, indexed by a finite set $\mathcal{M}$. Each asset $j \in \mathcal{M}$ is characterized by a value per share $v_j \in \mathbb{R}{\geq}\, 0$, assembled into the vector $\boldsymbol{v} \in \mathbb{R}^m$. Bank i holds $s_{ij} \in \mathbb{R}$ shares of asset $j \in \mathcal{M}$,

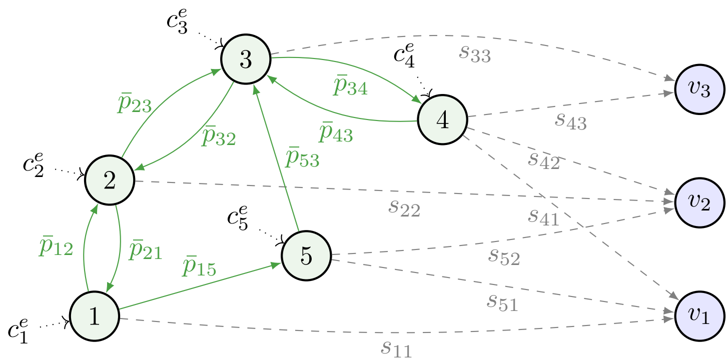

Figure 1: Illustration of a financial network with interbank network $\mathcal{G}$ (solid green edges) and external asset network $\mathcal{H}$ (dashed gray edges).

where strictly positive and negative values represent long and short positions, respectively. These entries s_{ij} are arranged into an $n \times m$ matrix $S \in \mathbb{R}^{n \times m}$.

Such an interconnection structure can also be represented as a graph, specifically, a weighted directed bipartite graph $\mathcal{H} = (\mathcal{V} \cup \mathcal{M}, \mathcal{F}, S)$, where a directed edge $(i, j) \in \mathcal{F} \subseteq \mathcal{V} \times \mathcal{M}$ exists if and only if $s_{ij} \neq 0$. We refer to this as the *external asset network* (depicted in gray in Fig. 1).

In summary, the external net position for bank $i \in \mathcal{V}$ can be written as $c_i = c_i^e + \sum_{j \in \mathcal{M}} s_{ij} v_j$, which we can write in compact form as

$$\boldsymbol{c} = \boldsymbol{c}^{\boldsymbol{e}} + S\boldsymbol{v}. \tag{3}$$

Note that the Eisenberg-Noe clearing mechanism can be extended to financial networks with cross-holdings (Elsinger, 2009), where banks can own shares in one another. Beyond the added complexity of defining a bank's equity as a function of its assets, the properties of the clearing vector (which will be introduced in Section 2.3) differ significantly from those in the original model (Eisenberg & Noe, 2001; Elsinger et al., 2006). In particular, the clearing vector is not a concave function of the vector $\boldsymbol{c}$ (Elsinger, 2009), and its computation does not reduce to a convex optimization problem like one presented in Proposition 1. Furthermore, as discussed in Elliott et al. (2014), a realistic model of banks' equities in the presence of cross-holdings should account for nonlinear failure costs, resulting in nonlinear systems of equations with multiple solutions, even when banks have no claims on each other ($\bar{p}_{ij} = 0$). Since our resilience analysis relies on characterizing clearing vectors and system loss through linear

Table 1: Model variables and parameters.

$\mathcal{V}$	set of banks
$\mathcal{M}$	set of external assets
c_i	external net position of bank $i \in \mathcal{V}$
$\bar{p}_{ij}$	liability from bank $i \in \mathcal{V}$ to bank $j \in \mathcal{V}$
$\bar{p}_i$	total debt of bank $i \in \mathcal{V}$
p_{ij}	actual clearing payment bank $i \in \mathcal{V}$ to bank $j \in \mathcal{V}$
c_i^e	net liquidity of bank $i \in \mathcal{V}$
s_{ij}	shares of asset $j \in \mathcal{M}$ owned by bank $i \in \mathcal{V}$
v_j	value per share of asset $j \in \mathcal{M}$

programming, we exclude cross-holding structures from this study, leaving them for future research.

We refer to a *financial network* as the combination of an interbank network and an external asset network, potentially including net liquidity (Fig. 1). The variables and parameters of our model are summarized in Table 1.

2.3. Clearing Vectors and System Loss

In regular operations, each bank $i \in \mathcal{V}$ has a nominal external net position $c_i = \bar{c}_i$ sufficiently large so that the bank is able to pay its debts at the end of the period, possibly by liquidating some of its external assets, in other words, the nominal equity of the bank, defined as the sum of the nominal external net position and the credits to other banks minus its liabilities, is nonnegative:

$$\bar{w}_i := \bar{c}_i + \sum\nolimits_{k \in \mathcal{V}} \bar{p}_{ki} - \bar{p}_i \geq 0, \quad \forall i \in \mathcal{V}. \tag{4}$$

However, if a financial shock impacts the network, the actual external net position of some bank i may decrease from its nominal value to $c_i < \bar{c}_i$, e.g., due to a decrease in an external asset's value, as we shall see below. In such a scenario, bank i may be unable to meet its obligations even if its debtors fulfill their liabilities completely, as its actual equity defined as

$$w_i := c_i + \sum_{k \in \mathcal{V}} \bar{p}_{ki} - \bar{p}_i \tag{5}$$

appears to be negative. We say that bank i incurs a *primary default* because its inability to fulfill liabilities is caused directly by the external shock. In this case, bank i should fulfill its liabilities to the best of its capacity, prioritizing obligations to the external sector, followed by those to other banks. As a result, the *actual payment* p_{ij} from i to j, in general, may be less than its nominal value $\bar{p}_{ij}$, decreasing the asset side of bank j's balance sheet. This, in turn, may lead to *secondary defaults*: A bank j that could meet its obligation $\bar{p}_j$ if its debtors fulfilled their obligations in full will actually default if those obligations are not met and

$$c_j + \sum_{k \in \mathcal{V}} p_{kj} - \bar{p}_j < 0. \tag{6}$$

In the latter case, the actual claims of some other banks on j must be reduced from $\bar{p}_{jk}$ to $p_{jk} < \bar{p}_{jk}$, to tertiary defaults and further cascades. In this way, even a single shock can trigger multiple bank defaults, causing a substantial shortfall across the financial system.

Eisenberg & Noe (2001), further generalized in Elsinger et al. (2006), proposed a constructive method for determining the actual clearing payments $p_{ij} \leq \bar{p}_{ij}$ between banks in the event of one or more defaults. One assumption imposed in Eisenberg & Noe (2001) is the equal priority of debt claims, implying proportional sharing in case of default (the pro-rata division rule). Thus, if $p_i = \sum_{j \in \mathcal{V}} p_{ij}$ is the total clearing payment from bank i to other banks, then $p_{ij} = a_{ij} p_i$ for all j. In view of this rule, it suffices to determine the vector of clearing payments $\boldsymbol{p} \in \mathbb{R}^n$, composed of the values p_i. Two additional rules relate to the banks' actual equities after fulfilling the clearing payments and can be expressed as follows Elsinger et al. (2006):

$$p_i = \begin{cases} \bar{p}_i & \text{if } \; c_i + \sum_{k \neq i} a_{ki} p_k \geq \bar{p}_i, \\ c_i + \sum_{k \neq i} a_{ki} p_k & \text{if } \; \bar{p}_i > c_i + \sum_{k \neq i} a_{ki} p_k \geq 0, \\ 0 & \text{if } \; c_i + \sum_{k \neq i} a_{ki} p_k < 0. \end{cases} \tag{7}$$

These equations imply, on one hand, the *limited liability rule*: a bank's total payment must not exceed its net residual value, so that its equity remains nonnegative (except in cases of insolvency, where equity is negative even without interbank payments). On the other hand, they impose the *absolute debt priority rule*: bank owners receive no value unless all nominal liabilities are fully cleared. Rewriting these equations in compact form leads to the definition of a clearing vector (Elsinger et al., 2006).

Definition 1. *A clearing vector for the interbank network* $(\mathcal{V}, \mathcal{E}, \bar{P})$ *with vector* *of net external positions* $\boldsymbol{c}$ *is a solution* $\boldsymbol{p}$ *to the following equation*

$$\boldsymbol{p} = \min(\bar{\boldsymbol{p}}, (\boldsymbol{c} + A^\top \boldsymbol{p})^+). \tag{8}$$

It is known (this follows, e.g., from the Knaster-Tarski fixed point theorem (Eisenberg & Noe, 2001; Elsinger et al., 2006; Glasserman & Young, 2016)) that at least one clearing vector always exists. Furthermore, the set of clearing vectors contains both a least element $\boldsymbol{p}_*(\boldsymbol{c})$ and a greatest element $\boldsymbol{p}^*(\boldsymbol{c})$, that is, for every solution to Eq. (8) one has

$$\boldsymbol{p}_*(\boldsymbol{c}) \leq \boldsymbol{p} \leq \boldsymbol{p}^*(\boldsymbol{c}). \tag{9}$$

In the generic situation, the least and the greatest clearing vectors coincide $\boldsymbol{p}_*(\boldsymbol{c}) = \boldsymbol{p}^*(\boldsymbol{c})$, and the clearing vector is uniquely determined by $\boldsymbol{c}$. Necessary and sufficient criteria for uniqueness are provided in Csóka & Herings (2024); Calafiore et al. (2024) for the case where $\boldsymbol{c} \geq 0$ and in Hurd (2016); Massai et al. (2022) for the general case $\boldsymbol{c} \in \mathbb{R}^n$.

In exceptional cases where the uniqueness cannot be guaranteed, we will assume that the banks use the greatest clearing vector $\boldsymbol{p}^*(\boldsymbol{c})$, which provides the minimal value of *system loss* (Glasserman & Young, 2016) (the total shortfall across the financial network) $L(\boldsymbol{p})$ among all clearing vectors. Here,

$$L(\boldsymbol{p}) = \mathbf{1}^\top(\bar{\boldsymbol{p}} - \boldsymbol{p}) = \|\bar{\boldsymbol{p}} - \boldsymbol{p}\|_1 = \sum_i (\bar{p}_i - p_i). \tag{10}$$

For a given vector $\boldsymbol{c}$, the greatest clearing vector (and the system loss associated to it) can be determined, e.g., through iterations of a monotone operator (Calafiore et al., 2025) or a modified fictitious default algorithm (Rogers & Veraart, 2013). However, since we aim to evaluate the *worst-case* system loss over a class of unknown vectors $\boldsymbol{c}$, these characterizations appear to lack of relevance. Instead, we use an extremal characterization of the greatest clearing vector (Proposition 1), which, as discussed in APPENDIX A, applies to "moderate" shocks that do not lead to insolvencies. Clearly, if $\boldsymbol{c} \geq \bar{\boldsymbol{c}}$, then no defaults are present due to Eq. (4), and thus $\boldsymbol{p}^*(\boldsymbol{c}) = \bar{\boldsymbol{p}}$, and there is no loss in the financial system $L(\boldsymbol{p}^*(\boldsymbol{c})) = 0$.

2.4. Resilience margin and worst-case loss

In this paper, we focus on financial shocks arising from fluctuations in actual asset prices. Specifically, we assume that the actual price vector $\boldsymbol{v}$ in Eq. (3)

can fluctuate and is expressed as

$$\boldsymbol{v} = \bar{\boldsymbol{v}} + \boldsymbol{\delta}, \tag{11}$$

where $\boldsymbol{\delta} \in \mathbb{R}^m$ represents the vector of price fluctuations, and $\bar{\boldsymbol{v}}$ denotes the nominal value. For instance, in contexts where fluctuations might be captured by stochastic processes, the nominal value $\bar{\boldsymbol{v}}$ can be interpreted as the expected value of the price for share. The vector of actual external net positions is given by

$$\boldsymbol{c} = \bar{\boldsymbol{c}} + S\boldsymbol{\delta}, \tag{12}$$

where $\bar{\boldsymbol{c}} := \boldsymbol{c}^{\boldsymbol{e}} + S\bar{\boldsymbol{v}}$ is the vector of nominal external net positions, and $S\boldsymbol{\delta}$ captures the impact of price fluctuations. In the absence of fluctuations $\boldsymbol{\delta} = \boldsymbol{0}$, the condition from Eq. (4) is assumed to hold, which is reformulated as follows.

Assumption 1. *The quantities A, $\bar{\boldsymbol{p}}$, S, $\boldsymbol{c}^{\boldsymbol{e}}$, $\bar{\boldsymbol{v}}$ are such that* $\boldsymbol{c}^{\boldsymbol{e}} + S\bar{\boldsymbol{v}} + A^\top \bar{\boldsymbol{p}} - \bar{\boldsymbol{p}} \geq 0.$

In presence of price fluctuations, as has been discussed in the previous section, a natural measure of the loss in the financial system is given by

$$\eta(\boldsymbol{\delta}) := L(\boldsymbol{p}^*(\bar{\boldsymbol{c}} + S\boldsymbol{\delta})), \tag{13}$$

which represents the loss corresponding to the greatest clearing vector compatible with the vector of the banks' actual external net positions $\boldsymbol{c} = \bar{\boldsymbol{c}} + S\boldsymbol{\delta}$. While this loss can be computed for each fluctuation vector $\boldsymbol{\delta}$, the price fluctuations in practice are unpredictable. This uncertainty leads to the problem of evaluating the *worst-case loss*, under the assumption that the magnitude of the fluctuation vector $\boldsymbol{\delta}$ is bounded in some norm: $\|\boldsymbol{\delta}\| \leq \varepsilon$. Different norms can be used to study different aspects of the price fluctuation. For instance, the ℓ_∞ norm ($\|\boldsymbol{\delta}\|_\infty = \max_{i \in \mathcal{M}} |\delta_i|$) captures independent variation of each asset price, while a bound in the ℓ_1 norm ($\|\boldsymbol{\delta}\|_1 = \sum_{i \in \mathcal{M}} |\delta_i|$) poses a constraint on the sum of the amplitudes of all perturbations. In other words, the worst-case loss is defined as the maximum value in the optimization problem

$$\begin{aligned} \eta_{wc}(\varepsilon) = \max\nolimits_{\delta} \quad & \eta(\boldsymbol{\delta}) \\ \text{s.t.:} \quad & \|\boldsymbol{\delta}\| \leq \varepsilon, \end{aligned} \tag{14}$$

where $\eta(\boldsymbol{\delta})$ is defined in Eq. (13). It can be shown (Calafiore et al., 2025) that $\eta(\boldsymbol{\delta})$ is a *convex* function, making its *maximization* over a convex set a nontrivial problem. However, this problem can be addressed for the ℓ_∞ and ℓ_1

norms, as described above, assuming that ε is not too large. As demonstrated in Calafiore et al. (2025), under these conditions, the worst-case loss computation reduces to a max-min problem that can be solved using LP duality. The method from Calafiore et al. (2025) enables the determination of the following financial network's characteristics.

First, one can determine the *resilience* margin (Calafiore et al., 2025), i.e., the maximum fluctuation magnitude ε^* such that no fluctuation vector $\boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\| \leq \varepsilon^*$ leads to defaults. In other words, for $\boldsymbol{c} = \bar{\boldsymbol{c}} + S\boldsymbol{\delta}$, banks can fully meet their debt obligations while keeping all equities from Eq. (5) nonnegative.

Second, the method from Calafiore et al. (2025) determines the maximum level of admissible fluctuation ε_{ub} such that price fluctuations with $\|\boldsymbol{\delta}\| \leq \varepsilon_{ub}$ cause a shock that does not result in insolvencies. More formally, for the vector of external net positions $\boldsymbol{c} = \bar{\boldsymbol{c}} + S\boldsymbol{\delta}$, a clearing vector exists under which all liabilities to the external sector can be fully cleared, even if defaults occur with respect to interbank liabilities. The optimal, in the sense of the system loss, clearing vector is the greatest clearing vector $\boldsymbol{p}^*(\boldsymbol{c})$, which, in this case, can be found from an LP (see Proposition 1). The number ε_{ub}, called the *insolvency margin*, can also be found from a similar linear program (Calafiore et al., 2025, Theorem 6).

Finally, for each $\varepsilon \in [\varepsilon^*, \varepsilon_{ub}]$, it is possible to compute the worst-case system loss $\eta_{wc}(\varepsilon)$ for fluctuations satisfying $\|\boldsymbol{\delta}\| \leq \varepsilon$, under the ℓ_1 and ℓ_∞ norms. In these two important cases, the optimization problem in Eq. (14) reduces to a single linear program (for the ℓ_1 norm) or a set of independent linear programs (for the ℓ_∞ norm), as detailed in Propositions 2 and 3). These results form the basis for our numerical study.

2.5. Research problem

In the remainder of this paper, we build on the computational results summarized above to conduct an extensive numerical campaign exploring how the structure of the interbank network and the external asset network influence the resilience of the financial network as a whole. Specifically, in the first set of studies, we focus on the interbank network $\mathcal{G}$ and examine key features such as connectivity, clustering, and degree distribution. These results expand some of the findings preliminarily analyzed in Zino et al. (2025), where a simpler scenario is studied in the absence of an explicit external asset network. In the second set of studies, our analysis shifts to the external asset network $\mathcal{H}$. In order to guarantee that Assumption 1 is satisfied, for each institution $i \in \mathcal{V}$, we assume that

$$\bar{c}_i = \bar{p}_i - \sum_{k \in \mathcal{V}} \bar{p}_{ki} + \xi_i, \tag{15}$$

where ξ_i is a realization of a random variable uniformly distributed in $[0, 100]$, with each realization independent of the others. Hence, in the absence of fluctuations no defaults are possible, and all institutions have the same expected nominal equity $\bar{w}_i$ from Eq. (4), with different realizations due to stochasticity.

We use the tools from Propositions 2 and 3 to compute the worst-case loss (η_{wc}) for increasing values of the maximal norm of the admissible fluctuation not exceeding the insolvency margin, $\varepsilon \leq \varepsilon_{ub}$. To achieve this, we generate the worst-case loss curve by solving linear programs at different (evenly spaced) levels of fluctuation within the range $[0, \varepsilon_{ub}]$.

In each study, we generate financial networks (i.e., $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \bar{P})$ and $\mathcal{H} = (\mathcal{V} \cup \mathcal{M}, \mathcal{F}, S)$) with the desired characteristics. In order to obtain statistically significant results we introduce stochasticity in the network formation process and we average the simulation outcomes over each random scenario. Specifically, in each of our studies we generate 100 independent instances of each network, and all the results presented hereafter are obtained by averaging the outputs of these 100 independent instances. All simulations are performed using MATLAB, and the code is available at https://github.com/lzino90/financial.

3. RESULTS

This section provides the details of our numerical experiments, together with our numerical results.

3.1. Impact of interbank network

In this section, we present and discuss the results of our numerical investigation of the role of the interbank network on the resilience of a financial system and the worst-case loss in the presence of defaults. To perform such a study, we consider a network of $n = 200$ banks. The interbank network is generated depending on the specific feature under analysis, as detailed below. The external asset network is instead kept fixed throughout all the numerical experiments in this section, and it comprises $m = 4$ external assets. Each bank $i \in \mathcal{V}$ has shares in either 1, 2, 3, or 4 assets, with uniform probability across all possible configurations. Then, the number of shares in each asset is sampled uniformly at random and the resulting values are normalized, so that $s_{i1}+s_{i2}+s_{i3}+s_{i4} = 100$ for all $i \in \mathcal{V}$.

3.1.1. Network connectivity

We start by investigating the impact of the network connectivity, viz. of the density of the interbank network. As discussed in the introduction, the impact of such a feature is complex, as it apparently leads to two contrasting effects, whereby the resilience could be either increased (Allen & Gale, 2000; Freixas et al., 2000; Babus, 2016), or decreased (Glasserman & Young, 2016; Acemoglu et al., 2015; Nier et al., 2007; Haldane & May, 2011; Bardoscia et al., 2017).

To investigate such a problem, we proceed as follows. We assume that the network $\mathcal{G}$ is a (weighted) directed regular random graph with out-degree equal to d (i.e., each bank has liabilities to exactly d other banks). We explore the role of the network connectivity by computing the worst-case loss as a function of the norm of the fluctuation ε (computed using our theoretical tools), for different values of $d \in \{1, \ldots, 10\}$. Clearly, larger values of d model networks with a denser connectivity pattern. Specifically, for each value of d, we generate a directed regular random graph of out-degree d. Such a network is generated as follows. For each node $i \in \mathcal{V}$, we sample a d-uple of nodes uniformly at random in $\mathcal{V} \setminus \{i\}$, with each institution independent of the others. Then, i is connected to these nodes. The total liability is split evenly across the (out)-neighbors, with each entry equal to $100/d$, so that $\bar{p}_i = 100$ for all $i \in \mathcal{V}$. Note that, in general, we obtain a directed network, due to the independent mechanisms through which each node selects its (out-)neighbors.

The results of our numerical experiments are represented in Fig. 2. First, we observe that the results obtained with the two different norms are qualitatively consistent. Predictably, fluctuations bounded in the ℓ_1 norm (i.e., with bounds on the total sum of the fluctuations and not just on its maximal value) lead to smaller total worst-case losses than the one bounded in the ℓ_∞ norm. Interestingly, we observe that the network connectivity impacts the performance of the financial network. In fact, curves corresponding to very sparse networks ($d = 1$ or $d = 2$) are shorter (i.e., the insolvency margin ε_{ub} is smaller) and they are above the others. This means that such network structures have low resilience and typically lead to larger worst-case losses with respect to more

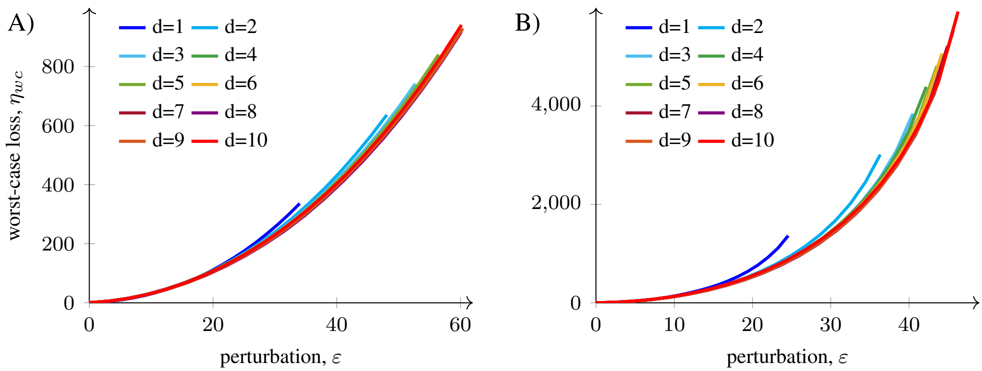

Figure 2: Impact of the network connectivity, computed numerically over 100 independent realizations of random regular graphs with different out-degree d, with fluctuations bounded in the A) ℓ_1 and B)ℓ_∞ norm.

connected structures.

However, it is worth noticing that the marginal benefit of increasing the network connectivity quickly decreases as the connectivity increases. In fact, for a network with $n = 200$ banks, one can still see some very minimal improvements up to $d = 5$; but then all curves when $d > 6$ are practically indistinguishable. Hence, we conclude that, while very poorly connected networks are weak and prone to default cascades, the beneficial effect of increasing the network connectivity seems to quickly decline.

3.1.2. Clustering

Many real-world financial systems are characterized by a phenomenon, whereby it is more likely that pair of entities that share common connections are, in turn, connected one to the other (Onnela et al., 2004; Rakib et al., 2021). In network science, such phenomenon is called clustering. Here, we investigate how clustering may impact the financial system. To this aim, we generate each instance of the interbank network as a Watts–Strogatz small-world network with average degree $d = 4$ (within these networks undirected, in- and out- degrees coincide), with a different re-wiring probability,[1] which allows us to generate

[1] Watts–Strogatz small-world networks are obtained starting from a regular lattice (which is highly clustered), and re-wiring edges in a stochastic fashion in order to add long-range interactions, which decrease the network clustering and the network diameter (Newman, 2018).

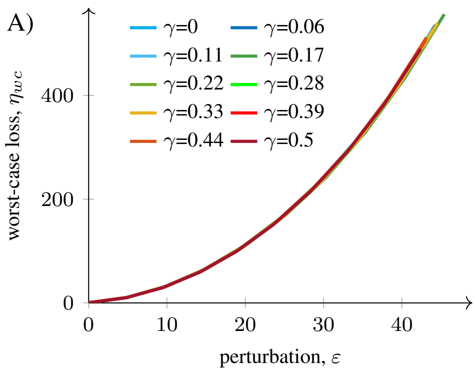

Figure 3: Impact of clustering, computed numerically over 100 independent realization of random regular graphs with different clustering coefficient γ, with fluctuations bounded in the A) ℓ_1 and B) ℓ_∞ norm.

networks with the same average degree, but with different levels of clustering. In fact, there is a one-to-one relation between the re-wiring probability and the network clustering coefficient, which is denoted by γ and represents the number of triadic closures in the network (in plain words, the larger γ, the higher the level of clustering). See, Newman (2018) for more details. Then, similar to the previous study, for each desired level of clustering we generate realizations of such a network structure and we set nonzero liabilities to $\bar{p}_{ij} = 25$. This makes the simulation comparable to the earlier scenarios where $\bar{p}_i = 100$.

Our numerical results, represented in Fig. 3, suggest that the level of clustering in the financial network has a very marginal impact on its performance. In particular, we observe that the worst-case loss seems not to be affected by the level of clustering. However, we observe that highly clustered networks tend to have slightly smaller values of the insolvency margin ε_{ub}. This means that highly clustered networks are slightly less resilient to large price fluctuations. This marginal impact may be due to the presence of two contrasting phenomena. On the one hand, once one or multiple defaults occur within a cluster of banks, all the other banks in that cluster have large exposure to them, possibly leading to further defaults. On the other hand, the other banks in the network would have little exposure to such banks, and are thus less likely to have defaults.

3.1.3. Heterogeneity

Another important feature of real-world financial networks is the presence of heterogeneity across different banks. To investigate the impact of heterogeneity, we compare the performances of the system on network structures with different levels of heterogeneity. Specifically, we consider random graphs with the following out-degree distributions: i) homogeneous, with each node having the same number of neighbors $d = 4$; ii) moderately heterogeneous, with each node having number of neighbors d_i sampled from a zero-truncated Poisson distribution (Ross, 2019) with mean 4; and iii) highly heterogeneous, with each node having number of neighbors d_i sampled from a Zipf distribution with mean equal to 4. All networks are then generated by sampling, for each node i, the set of (out-)neighbors as a d_i-uple of nodes uniformly at random, each node independent of the others. Similar to the other studies, we set the nonzero liabilities to $\bar{p}_{ij} = 100/d = 25$, to make this scenario comparable with the previous ones.

Our results, reported in Fig. 4, depict a nontrivial behavior. In fact, on the one hand, heterogeneity seems always detrimental to the overall network resilience reducing the value of the insolvency margin ε_{ub}, especially for fluctuations bounded in the ℓ_∞ norm. This means that it becomes easier for default cascades to lead to insolvency to the external sector when the system is heterogeneous. On the other hand, the impact on the worst-case loss is nonmonotonic. In fact, consistently across the two different norms, it seems that smaller levels of heterogeneity (green curve) might slightly decrease the worst-case loss, consistent with the results in other model-based studies (Caccioli et al., 2012). However, such a beneficial effect could be jeopardized by further increasing the heterogeneity (see, e.g., the red curve), consistent with the results in Glasserman & Young (2015).

We conjecture that a reason for such a nontrivial behavior can be that nodes with lower out-degree, which may be less robust in the face of price fluctuations, typically generate smaller cascades when they fail, yielding a smaller total loss. However, in highly heterogeneous networks, the default of some marginal nodes can ultimately lead one of the core banks to a default, yielding a larger cascade and, ultimately, a larger total loss. These results, which are consistent with results from other studies on the nontrivial and nonlinear impact of diversity and heterogeneity in the interbank network (Iori et al., 2006; Battiston et al., 2012), suggest that finding an optimal level of diversity in the network structure

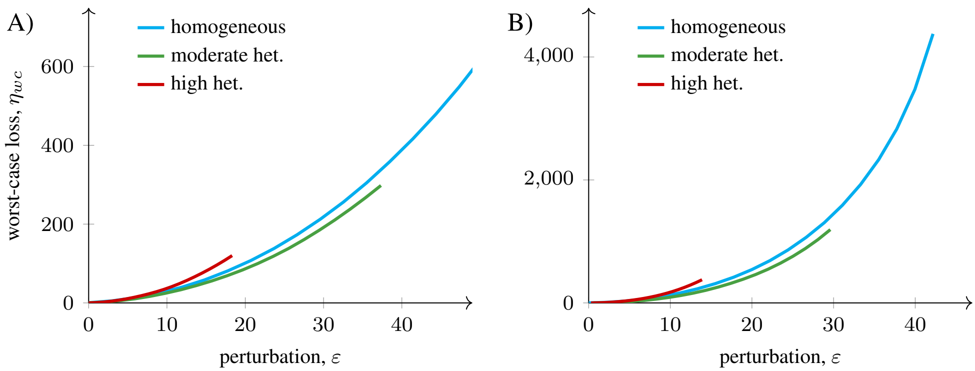

Figure 4: Impact of out-degree heterogeneity, computed numerically over 100 independent realizations of random regular graphs (homogeneous, cyan curve), Erdos-Renyi random networks (low heterogeneity, green curve) and Albert–Barabasi scale-free networks (high heterogeneity, red curve), with fluctuations bounded in the A) ℓ_1 and B) ℓ_∞ norm.

is indeed a key but nontrivial problem, which should be further investigated.

3.2. Impact of external asset network

Finally, we analyze how the characteristics of the external asset network impact the resilience of the financial network and its worst-case loss in case of price fluctuations. To perform such studies, we fix the characteristics of the interbank network across all numerical experiments to be a directed regular random graph with $n = 200$ banks (see Section 3.1.1 for the details on how such networks are generated), each one connected with other $d = 4$ banks with total liability $\bar{p}_i = 100$ for each $i \in \mathcal{V}$ split evenly across the 4 (out-)neighbors. Then, we generate external asset networks with different features, to assess their impact on the system. In all these networks, we consider $m = 8$ external assets. In this analysis, we consider only worst-case losses when considering fluctuations bounded in the ℓ_1 norm, while fluctuations bounded in the ℓ_∞ norm are omitted, being less interesting. In fact, ℓ_∞ norm captures the maximal fluctuation among all prices. This decouples the effect of the different assets, reducing the effect of the network.

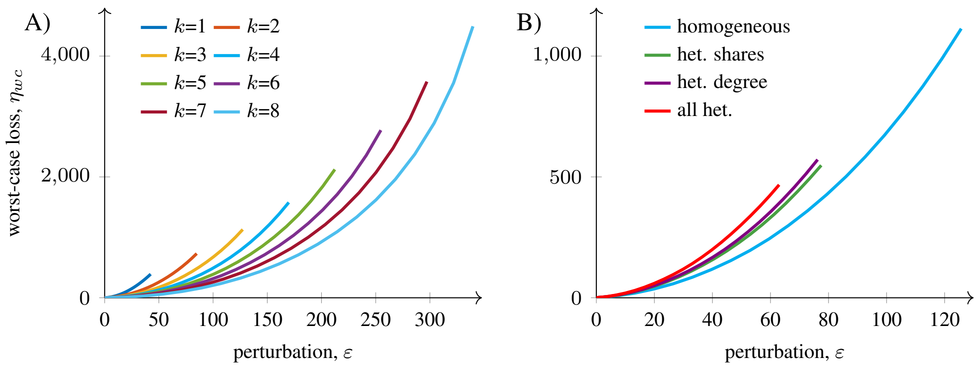

Figure 5: Impact of the external asset network, computed numerically over 100 independent realization of the financial network, with fluctuations bounded in the ℓ_1 norm. In A), we consider an external asset network with different connectivity (i.e., different degree k); in B), we compare the homogeneous scenario (cyan) with respect to heterogeneous scenarios in the number of shares across assets (green), the number of assets (violet), and both features (red).

3.2.1. Diversification

In our first study, we investigate the impact of diversification in the assets. To this aim, we consider scenarios in which each bank has shares in exactly k external assets (sampled as a k-uple for each bank, independent of the others), with a total amount of 100 shares divided evenly across the banks. In our analysis we vary the value of k in the discrete interval from 1 to $m = 8$ and we use the theoretical tools developed in Proposition 2 to compute the worst-case loss when considering fluctuations bounded in the ℓ_1 norm. Our results are reported in Fig. 5A).

Consistent with intuition and with the economic theory, we conclude that diversification is beneficial in financial networks. In fact, in scenarios in which each bank divides its shares among more external assets, the financial network becomes more resilient to larger fluctuations in price and, in correspondence of the same level of fluctuation, lower worst-case losses are obtained.

3.2.2. Heterogeneity

Finally, we investigate the role of diversity and heterogeneity in the external asset network. Specifically, we investigate such an effect along two directions, viz. by allowing banks to have i) a different number of shares in their assets and ii) shares in a different number of assets. The results of our analysis are reported in Fig. 5B).

Specifically, we compare a scenario in which each bank has shares in exactly $k = 3$ assets with 33.3 shares in each asset (cyan curve); a second scenario in which the total 100 shares are divided uniformly at random across the $k = 3$ assets (green curve); a third scenario in which each bank has 100 shares divided evenly across a random number of assets[2] (violet curve); and, finally, a fourth scenario in which both features of heterogeneity are included (red curve).

Our results suggest that, while diversification across the available assets is beneficial to the financial network, having heterogeneity in how banks allocate their shares (both in the number of assets and in the distribution of shares per asset) seems always disadvantageous. We conjecture that this is due to the inherent weakness of banks with poor diversification (either because they invest in few assets, or they allocate most of their shares in a single asset). These banks are thus more subject to a default, and may act as the weakest link in a chain, leading to default cascades in the entire financial network.

4. CONCLUSION

In this paper, we used an extended version of the Eisenberg–Noe model (Eisenberg & Noe, 2001; Elsinger et al., 2006) that accounts for external assets and, building on the theoretical developments in Calafiore et al. (2025), we investigated a critical problem in financial systems: *how does the structure of the financial network impact its resilience to default cascades?* To perform such a study, we systematically put forward a campaign of numerical experiments in which we assessed the resilience of the financial system under different assumptions regarding its network structure, and we computed the total worst-case loss in the presence of defaults due to fluctuations of the assets' prices.

The results of our numerical experiments suggest that both the network that describes the pattern of mutual liabilities between banks and the one describing how banks have shares in the different assets impact the performances of the financial network. In particular, we have determined that some features of the interbank network have a key role, such as connectivity and heterogeneity. In fact, very sparse networks are more likely to be subject to default cascades,

[2] Precisely, the number of assets is equal to a zero-truncated binomial random variable (Ross, 2019) with mean equal to 3.

whereas a moderate level of heterogeneity seems to be beneficial for decreasing worst-case losses in the presence of defaults. These observations corroborate similar results obtained with different methods and modeling approaches (Caccioli et al., 2012; Battiston et al., 2012; Iori et al., 2006). On the contrary, other features, such as the presence of clusters, seem to play only a marginal role in the emergent behavior of the financial network. Concerning the interactions of banks with assets, predictably, we have found that diversification is key to reducing the financial risk. On the contrary, heterogeneity in how banks invest in the external assets seems detrimental to the financial system, since it may generate some weak links in the network.

Besides elucidating the role of the network structure on the resilience of financial systems and on the worst-case loss in the presence of defaults, the results of our numerical experiments pave the way for several lines of future research. First, our results on the impact of heterogeneity in the interbank network depict a nontrivial scenario, whereby moderate levels of heterogeneity might be beneficial in reducing potential losses in financial networks. Further numerical studies should be pursued in order to determine the optimal level of heterogeneity and understand whether other features of the degree distribution may further impact the system's performances. Second, our numerical experiments rely on the theoretical tools developed in Calafiore et al. (2025). A future extension of the theoretical framework from Calafiore et al. (2025) to account for further features of financial systems, e.g., cross holdings between banks, the presence of illiquid assets, and defaults and bankruptcy costs (Cifuentes et al., 2005; Banerjee & Feinstein, 2022), would produce novel theoretical tools, which should be used to check the robustness of our findings and extend our numerical experiments. Moreover, the extension of the linear program framework above the insolvency margin, i.e., when fluctuations may lead to the insolvency of one or more banks, should be pursued by deriving novel theoretical results to compute the worst-case loss when a clearing vector does not exist. Third, the network features investigated in this paper are based on the characteristics empirically observed in financial systems (Onnela et al., 2004; Rakib et al., 2021). A relevant extension of this work is the analysis of case studies of real-world financial systems and the validation of our findings against the results obtained with other stress test methodologies (Amini et al., 2012; Battiston & Martinez-Jaramillo, 2018).

APPENDIX A. MATHEMATICAL METHODS

In this appendix, we summarize the mathematical framework established in Calafiore et al. (2025) to compute the maximum admissible fluctuation in the price of the asset before the financial network witnesses some defaults, and the worst-case loss, when fluctuations exceed the threshold.

A.1. Extremal Property of the Greatest Clearing Vector

The key step in characterizing the worst-case loss is outlined in the following proposition, which demonstrates that the minimal loss and the greatest clearing vector corresponding to the net external position vector $\boldsymbol{c}$ can often be determined through a simple linear program.

Proposition 1. *(Calafiore et al., 2025, Proposition 2) Consider the LP*

$$\begin{array}{rl} \min_{\boldsymbol{p}\in\mathbb{R}^n} & \mathbf{1}^\top(\bar{\boldsymbol{p}}-\boldsymbol{p}) \\ s.t.: & \mathbf{0}\leq\boldsymbol{p}\leq\bar{\boldsymbol{p}} \\ & \boldsymbol{c}+A^\top\boldsymbol{p}\geq\boldsymbol{p}. \end{array} \tag{16}$$

If the problem in Eq. (16) is feasible, then its optimal solution $\boldsymbol{p}^*$ is unique and coincides with the greatest clearing vector $\boldsymbol{p}^*(\boldsymbol{c})$. Moreover, under $\boldsymbol{p}^*$, all liabilities held by the external sector are paid. On the contrary, if the problem in Eq. (16) is infeasible, then there will exist banks which are insolvent with respect to the external sector for any admissible clearing vector $\boldsymbol{p}$.

When we consider the special case where $\boldsymbol{c} \geq 0$ (making the LP automatically feasible), we observe that this result appears in numerous works (Eisenberg & Noe, 2001; Glasserman & Young, 2016; Calafiore et al., 2024); the general case can be proved similarly.

The constraints in Eq. (16) ensure that the vector of interbank payments $\boldsymbol{p} \in \mathbb{R}^n$, under the proportionality rule $p_{ij} = p_i a_{ij}$, adheres to the limited liability principle: after clearing liabilities to the external sector and fulfilling these payments, the banks' equities remain nonnegative. As demonstrated in Calafiore et al. (2025), the feasibility of Eq. (16) is guaranteed for shocks whose magnitude does not exceed a critical threshold $\|\boldsymbol{\delta}\| \leq \varepsilon_{ub}$, which can be explicitly determined. The greatest clearing vector $\boldsymbol{p}^*(\boldsymbol{c})$ in the Eisenberg-Noe model is optimal among all such payment vectors, minimizing the system loss.

The infeasibility of Eq. (16) indicates that the net positions of the banks, $\boldsymbol{c}$, do not allow for clearing external liabilities without rendering at least one bank insolvent. This infeasibility can be addressed in several ways. One approach is to introduce additional binary variables to indicate whether a bank can remain solvent, as demonstrated in a similar extremal characterization of clearing vectors in Ararat & Meimanjan (2023, Corollary 4). Insolvent banks pay nothing to other banks and only partially fulfill their external liabilities. However, extending our results to MILP optimization presents a significant challenge. Moreover, defining system loss in the case of insolvencies constitutes a standalone research topic. Alongside the shortfall of banks, the loss to the external sector must be considered, as it can exacerbate the situation for banks in subsequent periods. Additionally, the costs associated with insolvencies must also be accounted for. On the other hand, by redefining the concept of system loss, one can consider other modifications to the LP in Proposition 1 to address potential infeasibility. For instance, the key constraint could be relaxed and replaced with a penalty term in the cost function

$$\begin{array}{rl} \min_{\boldsymbol{p}\in\mathbb{R}^n} & \mathbf{1}^\top(\bar{\boldsymbol{p}} - \boldsymbol{p}) + \boldsymbol{\mu}^\top(\boldsymbol{p} - \boldsymbol{c} - A^\top \boldsymbol{p})^+ \\ \text{s.t.:} & \mathbf{0} \le \boldsymbol{p} \le \bar{\boldsymbol{p}}. \end{array} \tag{17}$$

Here, $\boldsymbol{\mu}$ is a positive vector of penalty weights. Essentially, the "residual" $s_i = p_i - c_i - (A^\top \boldsymbol{p})_i$, when positive, can be interpreted as additional liquidity provided to a problematic bank by a regulating authority, similar to the model in (Calafiore et al., 2022). The penalty weights can significantly influence the optimum, and their selection requires separate analysis.

A.2. The Worst-Case Loss Computation

By applying the strong duality theorem in linear programming (Calafiore et al., 2025, Theorem 5), an explicit characterization of the worst-case loss can be derived in the case of ℓ_1 and ℓ_∞ norms of the asset price fluctuation vector.

Proposition 2. *(Calafiore et al., 2025, Section 4.1) For any level of fluctuation smaller than the insolvency margin $\varepsilon \le \varepsilon_{ub}$, the worst-case loss across all asset price variation vectors $\boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\|_1 \le \varepsilon$ is*

$$\begin{array}{rl} \eta_{wc} = \max_{i=1,\dots,m} \max_{\boldsymbol{\beta},\boldsymbol{\lambda}\ge 0} & (\mathbf{1} - \boldsymbol{\beta})^\top \bar{\boldsymbol{p}} - \bar{\boldsymbol{c}}^\top \boldsymbol{\lambda} + \varepsilon \boldsymbol{z_i}^\top \boldsymbol{\lambda} \\ \textit{s.t.:} & \boldsymbol{\beta} - \mathbf{1} + (I - A)\boldsymbol{\lambda} \ge 0, \end{array} \tag{18}$$

where $\boldsymbol{z_i}^\top$ is the component-wise absolute value of the i-th row of matrix S.

In other words, the worst-case loss can be efficiently computed by simply solving m independent linear programs, one associated with each asset.

Corollary 1. *The following properties concerning the set of arguments* $(i^*, \boldsymbol{\lambda}^*, \boldsymbol{\beta}^*)$ *that maximize the objective function in Eq.* (18) *hold true:*

1. *The argument i^* determines the asset which would have the largest impact on the financial network due to fluctuations in its value. More precisely, the vector $\boldsymbol{\delta}$ is defined as*

$$\delta_i = \begin{cases} 0 & \text{if } i \neq i^*, \\ -\varepsilon & \text{if } i = i^*. \end{cases} \tag{19}$$

2. *The argument $\boldsymbol{\lambda}^*$ identifies the set of banks that have a primary default in the worst-case scenario: $\lambda_j^* > 0$ if and only if j has a primary default.*

Proof. Item 1) has been proved in Calafiore et al. (2025). To prove item 2), from the constraint in Eq. (18) for the generic jth bank, we get $\beta_j + \lambda_j - 1 + (A\boldsymbol{\lambda})_j \geq 0$, where the last term depends on some entries of $\boldsymbol{\lambda}$, but not on λ_j as $a_{jj} = 0$. Hence, the constraint requires the sum $\beta_j + \lambda_j$ to be greater than some term independent of these two variables. Let us now focus on the contribution associated with β_j and λ_j to the objective function, which is equal to

$$(1 - \beta_j)\bar{p}_j + \lambda_j(\varepsilon s_{i^*j} - \bar{c}_j) = \bar{p}_j + \lambda_j(\varepsilon s_{i^*j} - \bar{c}_j) + \beta_j(-\bar{p}_j). \tag{20}$$

with the constraints that $\beta_j + \lambda_j \geq 1 + \sum_{k \neq j} a_{ij}\lambda_k$. Clearly, increasing β_j has always a negative impact. Moreover, keeping constant the sum of β_j and λ_j (which should be larger than $1 + \sum_{k \neq j} a_{jk}\lambda_k$), it is more convenient to have $\lambda_j > 0$ if and only if $\varepsilon s_{i^*j} - \bar{c}_j > \bar{p}_j$, i.e., if bank j has a primary default due to the perturbation $\boldsymbol{\delta}$ defined in Eq. (19). Hence,

$$\lambda_j^* = \begin{cases} 1 + \sum_{k \neq j} a_{jk}\lambda_k^* & \text{if } j \text{ has a primary default,} \\ 0 & \text{otherwise,} \end{cases} \tag{21}$$

which implies

$$\beta_j^* = \begin{cases} 0 & \text{if } j \text{ has a primary default,} \\ 1 + \sum_{k \neq j} a_{jk}\lambda_k^* & \text{otherwise,} \end{cases} \tag{22}$$

yielding the second claim. □

The case of ℓ_∞ norm leads to an even simpler formulation, where a single linear program is obtained, as detailed in the following.

Proposition 3. *(Calafiore et al., 2025, Section 4.2) For any level of fluctuation smaller than the insolvency margin $\varepsilon \leq \varepsilon_{ub}$, the worst-case loss across all price fluctuation vectors with $\|\boldsymbol{\delta}\|_\infty \leq \varepsilon$ is equal to*

$$\begin{array}{rl} \eta_{wc} = \max_{\boldsymbol{\beta},\boldsymbol{\lambda}\geq 0} & (\mathbf{1}-\boldsymbol{\beta})^\top \bar{\boldsymbol{p}} - \bar{\boldsymbol{c}}^\top \boldsymbol{\lambda} + \varepsilon \mathbf{1}^\top \tilde{S}^\top \boldsymbol{\lambda} \\ s.t.: & \boldsymbol{\beta} - \mathbf{1} + (I-A)\boldsymbol{\lambda} \geq 0, \end{array} \tag{23}$$

where $\tilde{S}$ is the component-wise absolute value of matrix S.

Remark 1. *Also for Proposition 3, the argument that maximizes the objective function in Eq. (23) provides information on the banks that are affected by such a fluctuation, as in Corollary 1.*

Remark 2. *For both Propositions 2 and 3, when the level of fluctuation is smaller than the resilience margin, i.e., $\varepsilon < \varepsilon^*$, the solution of the linear programs in Eq. (18) and Eq. (23) gives $\eta_{wc} = 0$, since no defaults occur.*

References

Acemoglu, D., Ozdaglar, A., & Tahbaz-Salehi, A. (2015). Systemic risk and stability in financial networks. *American Economic Review*, *105*(2), 564–608.

Allen, F., Babus, A., & Carletti, E. (2012). Asset commonality, debt maturity and systemic risk. *Journal of Financial Economics*, *104*(3), 519–534.

Allen, F., & Gale, D. (2000). Financial contagion. *Journal of Political Economy*, *108*(1), 1–33.

Amini, H., Cont, R., & Minca, A. (2012). Stress testing the resilience of financial networks. *International Journal of Theoretical and Applied Finance*, *15*(01), 1250006.

Amini, H., & Feinstein, Z. (2023). Optimal network compression. *European Journal of Operational Research*, *306*(3), 1439–1455.

Ararat, C., & Meimanjan, N. (2023). Computation of systemic risk measures: A mixed-integer programming approach. *Operations Research*, *71*(6), 2130–2145.

Babus, A. (2016). The formation of financial networks. *RAND Journal of Economics*, *47*(2), 239–272.

Banerjee, T., & Feinstein, Z. (2022). Pricing of debt and equity in a financial network with comonotonic endowments. *Operations Research*, *70*(4), 2085–2100.

Bardoscia, M., Battiston, S., Caccioli, F., & Caldarelli, G. (2017). Pathways towards instability in financial networks. *Nature Communications*, *8*, 14416.

Battiston, S., Delli Gatti, D., Gallegati, M., Greenwald, B., & Stiglitz, J. E. (2012). Liaisons dangereuses: Increasing connectivity, risk sharing, and systemic risk. *Journal of Economic Dynamics and Control*, *36*(8), 1121–1141.

Battiston, S., & Martinez-Jaramillo, S. (2018). Financial networks and stress testing: Challenges and new research avenues for systemic risk analysis and financial stability implications. *Journal of Financial Stability*, *35*, 6–16.

Caccioli, F., Catanach, T. A., & Farmer, J. D. (2012). Heterogeneity, correlations and financial contagion. *Advances in Complex Systems*, *15*(supp02), 1250058.

Calafiore, G., Fracastoro, G., & Proskurnikov, A. (2022). Control of dynamic financial networks. *IEEE Control Systems Letters*, *6*, 3206–3211.

Calafiore, G., Fracastoro, G., & Proskurnikov, A. (2024). Optimal clearing payments in a financial contagion model. *SIAM Journal on Financial Mathematics*, *15*(2), 473-502.

Calafiore, G., Fracastoro, G., & Proskurnikov, A. V. (2025). Default robustness and worst-case losses in financial networks. *Applied Network Science*.

Cifuentes, R., Ferrucci, G., & Shin, H. S. (2005). Liquidity risk and contagion. *Journal of the European Economic Association*, *3*(2-3), 556–566.

Csóka, P., & Herings, P. J.-J. (2024). Uniqueness of clearing payment matrices in financial networks. *Mathematics of Operations Research*, *49*(1), 232–250.

Eisenberg, L., & Noe, T. H. (2001). Systemic risk in financial systems. *Management Science*, *47*(2), 236–249.

Elliott, M., Golub, B., & Jackson, M. O. (2014). Financial networks and contagion. *American Economic Review*, *104*(10), 3115–53.

Elsinger, H. (2009). *Financial networks, cross holdings, and limited liability*. Oesterreichische Nationalbank Austria.

Elsinger, H., Lehar, A., & Summer, M. (2006). Risk assessment for banking systems. *Management Science*, *52*(9), 1301-1314.

Feinstein, Z. (2017). Financial contagion and asset liquidation strategies. *Operations Research Letters*, *45*(2), 109-114.

Freixas, X., Parigi, B., & Rochet, J.-C. (2000). Systemic risk, interbank relations, and liquidity provision by the central bank. *Journal of Money, Credit, and Banking*, *32*(3), 611–638.

Glasserman, P., & Young, H. P. (2015). How likely is contagion in financial networks? *Journal of Banking & Finance*, *50*, 383–399.

Glasserman, P., & Young, H. P. (2016). Contagion in financial networks. *Journal of Economic Literature*, *54*(3), 779–831.

Haldane, A., & May, R. (2011). Systemic risk in banking ecosystems. *Nature*, *469*,

351–355.
Hurd, T. (2016). *Contagion! Systemic Risk in Financial Networks*. Springer.
Iori, G., Jafarey, S., & Padilla, F. G. (2006). Systemic risk on the interbank market. *Journal of Economic Behavior & Organization*, *61*(4), 525–542.
Kusnetsov, M., & Veraart, L. A. M. (2019). Interbank clearing in financial networks with multiple maturities. *SIAM Journal on Financial Mathematics*, *10*(1), 37-67.
Massai, L., Como, G., & Fagnani, F. (2022). Equilibria and systemic risk in saturated networks. *Mathematics of Operation Research*, *47*(3), 1707-2545.
Newman, M. (2018). *Networks* (2nd ed.). Oxford, UK: Oxford University Press.
Nier, E., Yang, J., Yorulmazer, T., & Alentorn, A. (2007). Network models and financial stability. *Journal of Economic Dynamics and Control*, *31*(6), 2033–2060.
Onnela, J.-P., Kaski, K., & Kertész, J. (2004). Clustering and information in correlation based financial networks. *The European Physical Journal B*, *38*(2), 353–362.
Rakib, M. I., Nobi, A., & Lee, J. W. (2021). Structure and dynamics of financial networks by feature ranking method. *Scientific Reports*, *11*(1).
Rogers, L. C., & Veraart, L. A. (2013). Failure and rescue in an interbank network. *Management Science*, *59*(4), 882–898.
Ross, S. M. (2019). *Introduction to Probability Models* (12th ed.). San Diego CA, US: Academic Press.
Zino, L., Proskurnikov, A. V., Fracastoro, G., & Calafiore, G. C. (2025). Structure and shock resistance in financial networks: A numerical study. In *Complex Networks & Their Applications XIII* (p. 103–114). Springer Nature Switzerland.

Journal of Mechanism and Institution Design
ISSN: 2399-844X (Print), 2399-8458 (Online)
DOI:10.22574/jmid.2025.12.005

A POSITION ASSIGNMENT EXPERIMENT AMONG ACTIVE DUTY UNITED STATES NAVY PHYSICIANS

Richard Childers, University of California, San Diego, USA
rchilders@health.ucsd.edu

Alicea Mingo, Bureau of Medicine and Surgery, United States Navy, USA
aliceamingo@yahoo.com

Joel Schofer, US Navy Medicine Readiness and Training Command Guam
Captain, Medical Corps, United States Navy
jschofer@gmail.com

Naomi Utgoff, United States Naval Academy, USA
utgoff@usna.edu

William Howard Beasley, University of Oklahoma Health Campus, USA
william-beasley@ou.edu

ABSTRACT

The United States Navy (USN) assigns personnel to jobs via assignments officers; placement guidance exists but the process is decentralized. Participant attitudes towards and limited experience with market design are possible impediments to its incorporation into the assignment process. We investigate the appetite for the use of the deferred acceptance during the assignment process. We implemented a pilot project incorporating deferred acceptance into 231 assignments of 174 USN physicians to 23 commands. Pre-intervention, we

The views expressed in this article reflect the results of research conducted by the author(s) and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the United States Government. All authors with the exception of Naomi Utgoff were military service members during this project. This work was prepared as part of their official duties. Title 17 U.S.C. 105 provides that "Copyright protection under this title is not available for any work of the United States Government." Title 17 U.S.C. 101 defines a United States Government work as a work prepared by a military service member or employee of the United States Government as part of that person's official duties. The study protocol was approved by the Naval Medical Center San Diego Institutional Review Board in

surveyed physicians to obtain baseline perceptions of the traditional assignment process; post-intervention we surveyed pilot project participants. Only 38.2% of pre-pilot survey respondents said they preferred assignment via deferred acceptance; in the post-pilot survey 78.7% of participating physicians and 87.8% of participating commands reported a desire to keep using deferred acceptance. Our study suggests that attitudes towards using market design in personnel assignment may improve with experience, facilitating broader acceptance of its usage in the future.

Keywords: Market design, military personnel assignment, experiments.

JEL Classification Numbers: C78, C93, D47, M51.

1. INTRODUCTION

Military personnel assignment is a *de facto* matching market: service members have preferences over commands, commands have preferences over service members, and participants on one side of the market seek to match with participants on the other side. DoD, with over 1 million active duty personnel, represents an enormous market design opportunity. The United States Army has applied formal market design to aspects of the military personnel assignment (see for example Sönmez & Switzer (2013); Greenberg et al. (2024); Davis et al. (2023)) but the process remains largely decentralized and informal across DoD. General placement guidance is available, but there is no formalized decision-making process.

A potential impediment to the creation and implementation of a formal market design for military personnel assignment is participant attitudes towards such a change. This paper investigates the evolution of attitudes among

compliance with all applicable Federal regulations governing the protection of human subjects. Research data were derived from an approved Naval Medical Center San Diego IRB protocol. The authors declare there are no conflicts of interest. We thank Rear Adm. J.F. Meier, Mr. Philip Roth, CAPT Nanette Rollene, and CAPT Brendon Drew for their help completing this project. We thank John Hatfield for advice at the early stages of the project and comments on a draft. Naomi Utgoff gratefully acknowledges funding from the George Miller Trust as a USNA Miller Fellow during July–August 2022 and July–August 2023. Nayoung Rim (USNA) and Scott Kaplan (USNA) generously helped us with additional regression analyses during the revision of this paper. We are grateful to a reviewer whose feedback greatly improved the paper. Any errors are our own.

commands and USN physicians towards the use of market design in personnel assignment. Between 2017 and 2020, we implemented a pilot project incorporating deferred acceptance into 231 assignments of 174 USN physician assignments to 23 commands. Prior to the pilot project, we surveyed all 3,825 USN physicians to gauge their satisfaction with the existing decentralized process and attitudes towards the use of deferred acceptance in personnel assignment; we received 1,015 responses representing 951 usable responses. In 2018, 2019, and 2020, we surveyed the pilot project participants (23 commands and 152 USN physicians representing 185 assignments) and asked explicitly whether they preferred to continue to use deferred acceptance as part of the assignment process before and after the project to gauge attitudes towards the pilot project in particular and use of market design methods in general. (Unfortunately, we did not survey pilot project participants in 2017.) In our pre-pilot survey, only 38.2% of respondents reported that they would prefer to receive their next billet via deferred acceptance rather than via the existing decentralized USN assignment process. Participants in the pilot project subsequently reported significantly favorable attitudes towards the continued use of deferred acceptance in future personnel assignment: 87.8% of commands and 78.7% of USN physicians reported that they preferred to continue the use of deferred acceptance rather than return to the status quo decentralized process. Although our study is small, this result suggests that participant attitudes towards the use of market design in personnel assignment are flexible and may improve with exposure.

There is a growing literature on market design for military personnel assignment. Ferguson et al. (2020) use linear programming to find balanced personnel assignments between U.S. Army personnel and billets. They also minimize the number of blocking pairs present in an assignment, since military personnel assignment does not always admit a fully stable outcome. Korkmaz et al. (2008) propose a decision support system to assist detailers in making personnel assignments. (Yang et al., 2003) offer an algorithm which colocates dual military couples. Notably, the U.S Army implemented the Officer Career Satisfaction Program (Sönmez & Switzer, 2013) in 2006, and has more recently implemented a minimal market design for initial placement at both West Point and ROTC. (Greenberg et al., 2024). Davis et al. (2023) conducted an experiment using deferred acceptance to assign U.S. Army Officers and showed that while use of deferred acceptance decreases officer attrition in the short run, deferred acceptance was of limited use in improving performance evaluations or promotions.

Theoretically, military personnel assignment is a one-to-one matching problem with minimums. Gale & Shapley (1962) proposed the deferred acceptance algorithm which addresses one-to-one matching with no minimums. They showed that this algorithm terminates in polynomial time and its outcome is always stable, i.e., there are no mutual agreeable changes to the assignment. However, their algorithm does not (and cannot) accommodate the USN requirement that certain billets that must be filled with an active duty service member, and that each service member must be assigned to a billet. Indeed, Roth's Rural Hospital Theorem (Roth, 1986) says that the requirement that certain billets must be filled may preclude the existence of any stable assignment. The market design literature does not to our knowledge currently contain a suitable matching market design for USN's personnel assignment problem. Fragiadakis et al. (2016) offer the best known solutions in a setting in which all physicians are acceptable to all billets; their solutions remain approximate in USN's setting, since a physician's acceptability to certain billets depends on (among other things) the physician's most recent assignment.

The structure of this paper is as follows. Section 2 discusses Navy physician assignment, including institutional constraints and the existing assignment methodology. Section 3 describes the deferred acceptance experiment (*pilot project*) and the surveys conducted before and after the pilot project (*pre-pilot survey* and *post-pilot survey*) respectively. Section 4 analyzes the survey data and discusses its implications. Section 5 concludes. We use some terminology particular to USN; we define all terms in the text at first use and provide a glossary at the end of the paper.

2. NAVY PHYSICIAN ASSIGNMENT

USN employs approximately 4,000 physicians serving on active duty in various commands throughout the world. In general, these physicians rotate to new positions (henceforth *billets*) every three to five years. The process of assigning physicians to billets is complex; due to needs of USN, not every physician can be assigned their preferred billet. Additionally, USN must prioritize certain billets such as those in direct support of ship units. Billets involving solely clinic or hospital-based practices are lower priority and more easily supported by civilian physicians.

Billets differ across geographic location and assignment type. Geographically, USN categorizes billets as in the continental United States (CONUS) or outside the continental United States (OCONUS). There are two assignment types: *ship billets* (also known as operational) and *shore billets* (also known as non-operational). Physicians in ship billets work in direct support of military units, such as ships, infantry, special forces, or other uniquely military commands. Physicians in shore billets work at hospitals and clinics, or in administrative positions. We refer to commands featuring ship (operational) billets as *ship commands* and we refer to commands featuring shore (non-operational) billets as *shore commands*.

The desirability of these billets varies by physician but there are some trends. For example, billets in Southern California are generally more popular than those overseas or in isolated duty stations in the continental United States. In general, CONUS billets are more popular than OCONUS billets, and shore billets are more popular than ship billets. USN physicians can (and periodically must) rotate between geographic locations and ship and shore billets.

2.1. Institutional Constraints

Navy physician assignment features several constraints similar to the rest of the United States Military. Physicians can only fill billets they are qualified for based on specialty, officer rate, and physical fitness. Navy personnel cannot be forced into consecutive overseas tours, and their dependents must also qualify for certain billets. For example, certain locations may not have adequate medical or social support for a dependent's needs. In the event that the demand for USN physicians exceeds the supply, DoD may compel service member(s) to staff a higher priority billet(s), while staffing lower priority billets with civilians or else leaving them unfilled. All physicians must be assigned to a billet; like all service members, USN physicians cannot unilaterally exit USN.

2.2. Existing Assignment Methodology

The existing detailing process which determines the assignment of USN personnel to billets does not utilize any algorithm or other centralized market clearing process; rather, it is governed by a collection of policies delineated in the Navy Military Personnel Manual (MILPERSMAN) Articles 1300-100 – 1301-907 (USN, 2024). These articles appear on pages 1291–1593 of MILPERSMAN, with specific details relevant to physician career paths within

USN in Article 1301-815 on pages 1535–1538. In particular, Article 1301-100 outlines three criteria (in order of importance) which guide the detailing process:

1. Needs of the Navy
2. Career Needs of the Individual
3. Desires of the Individual

In the detailing process, *detailers* (also known as *assignment officers*) represent Naval personnel and *placement officers* represent the interests of the commands under their responsibility. There are multiple detailers (each of whom is responsible for a set of officers) and multiple placement officers (each of whom is responsible for a set of commands). Detailers help physicians gather information about billets that will be available for a physician's next assignment. Placement officers provide a list of hard requirements for each billet to the detailers. Detailers consider a physician's individual skill and seniority, individual requests, command needs, and mission priority among other factors when making billet assignments. Detailers and placement officers can and do communicate during the process in an effort to balance individual desires with the needs of the commands. General placement guidance is available to detailers, but there is not a formalized decision-making process. Detailers generally take each individual physician's preferences into account; the commands on the other side of this matching process often have very little input. Henceforth, we refer to assignment by detailers as the *status quo assignment process*.

2.3. The Deferred Acceptance Algorithm

The pilot project used the deferred acceptance algorithm (Gale & Shapley, 1962) to determine an initial assignment of USN physicians to billets, with manual adjustment of the algorithm's outcome to arrive at an assignment satisfying the institutional constraints discussed in Subsection 2.1. We state the physician-proposing deferred acceptance algorithm using language specific to USN physician assignment.

1. Step k:

   (a) Each physician proposes to his/her/their favorite billet that did not reject that physician in any previous step.

(b) Each billet holds its favorite physician among all of the physicians who proposed to that billet in Step k.

2. If there are no rejections, or if every rejected physician was rejected from his/her/their last choice billet, terminate the algorithm; each physician is assigned to the billet holding his/her/their proposal, or remains unassigned because he/she/they was rejected by all billets.

3. Otherwise, proceed to Step $k + 1$.

3. SURVEYS AND PILOT PROJECT

Our research consisted of an experimental implementation of deferred acceptance, with pre- and post-intervention surveys. The full text of all of the surveys appears in Appendix A.

1. First, we conducted a *pre-pilot survey* of all USN physicians to assess baseline perceptions about the status quo assignment process and assess attitudes towards the use of deferred acceptance. We surveyed all 3,825 USN physicians and received 1,015 responses of varying completeness.

2. Second, we executed a *pilot project* which incorporated the deferred acceptance algorithm into USN physician assignment. This project completed 231 assignments of 174 USN physicians to 23 commands between 2017 and 2020.

3. In 2018, 2019, and 2020, we conducted a *post-pilot survey* of pilot project participants (152 USN physicians and 23 commands representing 185 assignments) to assess the participants' attitudes towards the pilot project and continued use of deferred acceptance following the pilot project. Among physicians, our unit of observation was assignment to a billet; we received 184 responses to the post-pilot physician survey (response rate 99.5%) Among commands, our unit of observation was command-year; we received 49 of 62 responses to the post-pilot command survey (response rate 79.0%). Unfortunately, we did not complete a post-pilot survey in 2017.

3.1. Pre-Pilot Survey: Physician Attitudes

Investigators for this survey included one current physician detailer and two former physician detailers. The study population included all active duty physicians in USN who had received at least one billet assignment from 2012 or later. The analysis in Section 4.1 excludes physicians who reported not executing new orders after 2012 since such a long time had elapsed since they participated in the assignment process.

We built our survey using a government survey tool at https://max.gov. There is no validated survey instrument regarding this topic available, so the study authors created and trialed the questions on approximately fifteen USN physicians of varied experience. We adjusted the questions based on feedback from this process. The survey introduction contained a cover letter explaining the goals of the research, contact name and address of the primary investigator, and the aims of the study.

After endorsement by the USN Survey Authority, and approval as a performance improvement project from the Naval Medical Center San Diego Institutional Review Board, our survey was distributed between August 1, 2016 and October 31, 2016. The survey was distributed by email using physician specialty distribution lists. Each physician specialty has a *specialty leader* who acts as a liaison between the specialty and USN medicine. Each specialty leader sent the survey to individual members of that specialty. We did not give specific instructions to the specialty leaders about how to distribute the survey, and there was likely some variation among specialty leaders regarding the number of times it was sent out.

Survey participants directly reported their demographic information, including officer rate, medical specialty, current position, and most recent year in which the individual completed new orders. Individual responses were anonymous, but we compared demographic information of survey respondents at a group level to the demographic information of all USN physicians; USN maintains demographic information for all USN physicians in its personnel database known as the Officer Assignment Information System (OAIS).

We asked four questions to assess physicians' overall happiness with the current assignment process. Subjects were asked to rank, using a five point Likert scale with 1 being unsatisfied and 5 being very satisfied, their satisfaction with the status quo assignment process. We also asked them to rank their perceptions of transparency (favoritism) using a five point Likert scale with 1 being not transparent (a problem) and 5 being very transparent (not a problem at all). Finally, we asked respondents to label their current position as their first choice, second choice, etc. Perceptions of transparency and favoritism, and rank of the current position, correlated strongly with satisfaction. For simplicity, we only present results regarding satisfaction.

We determined *a priori* seven potential predictors of satisfaction with the status quo assignment process: officer rate, specialty type, staffing rate, bonus pay, deployment rate, preference for current orders and current billet category. To gauge preference for current orders, we elicited from survey respondents how they ranked their assigned billet: first choice, second choice, third choice, fourth choice, or fifth or lower choice. For clarity, we classified individual medical specialties into five cohorts: (a) surgical specialties, (b) non-surgical specialties, (c) family medicine, (d) residents (physicians who have completed medical school, but still require further specialty training), and (e) operational physicians. These cohorts are a traditional grouping of specialties, and they share similarities including types of available assignments. Note that operational specialties are not synonymous with assignment to a ship billet; rather, operational specialties are those specific to the practice of medicine in a military context.

As discussed in Section 2, USN physicians can and at times must rotate between geographic locations and between ship and shore billets. Our pre-pilot survey data includes responses from physicians across five billet types: CONUS shore, Administrative, CONUS ship, OCONUS shore, and OCONUS ship.

A deployment is an event in which a military member leaves the location where their home is based for three months or longer in support of operational forces. A common example involves one month of training at a CONUS location followed by seven months in the Middle East. Certain medical specialties deploy more frequently than others. There are no existing data to rank specialties based on their deployment rate; using significant anecdotal experience, the study authors identified the following specialties as having high deployment rates: anesthesia, emergency medicine, general surgery, orthopedic surgery, psychiatry, and pulmonary/critical care.

Military physicians are paid based on length of time in service and also receive *incentive pay*, the size of which varies by specialty. The size of this bonus generally reflects U.S. civilian physician salaries, so that specialists earn more than primary care physicians. We used the 2016 rate.

To facilitate comparisons to other military services, we used officer rate to describe survey participants. USN physicians enter service as an O3 (lieutenant) and based on performance may be promoted as high as O6 (captain). We refer to these rates as O3 (lieutenant or LT) , O4 (lieutenant commander or LCDR), O5 (commander or CDR), and O6 (captain or CAPT).

Staffing rates vary between specialties. A specialty's staffing rate is the number of USN physicians in that specialty divided by the number of billets in that specialty. If there are more USN physicians in a particular specialty than there are billets, the staffing rate is >100%. In general, USN physicians in understaffed specialties have less flexibility in assignment choice as they are needed to fill important billets. For example, USN may deny them an opportunity to fill a desirable position outside their specialty because of needs within their own specialty. Thus, staffing rate may affect satisfaction with the assignment system. We obtained staffing rates from USN's Bureau of Medicine and Surgery, the administrative lead for medical activities. We created a staffing average for the specialty using historical data from the years 2013–2016 as published by the Corps Chief's Office of USN Medicine.

3.2. Pilot Project: The Match

The pilot project was executed by detailers, including three of the authors, annually between 2017–2020 on the emergency medicine community. Participants consisted of (1) USN emergency medicine physicians and (2) command match representatives. The latter are the officers, one per command, responsible for submitting the command's ranked list of USN physicians to the pilot project. We distributed lists of billets to USN physicians and lists of USN physicians to relevant commands during the summer, one year prior to assignment availability. Participants had until December to gather information, interview, and generate their respective *rank lists* (preference relations) over the other side: each physician ranked the available billets starting with the physician's most preferred billet, and each command match representative listed the available physicians starting with the command's most preferred physician. We processed the reported rank lists using the deferred acceptance algorithm, the same algorithm used by the National Resident Matching Program (NRMP). We used a version derived from open-source software Klein (2018).

Following the execution of the deferred acceptance algorithm, some physicians remained unassigned and some billets remained empty. However, as noted in Section 2.1, an assignment in which a physician remains unassigned or in which one or more critical billets remain empty is not feasible for USN. Active duty personnel are under contract; USN needs to find them work, and it has billets that need to be filled. Thus, these physicians were manually matched. This process was called the scramble and was done by assignment personnel in a manner thought to follow Naval instruction. The scramble was complicated; due to military policy, we could not simply pair the unassigned physicians into empty billets due to the constraints discussed in Section 2.1. Thus, some individuals who were matched by the algorithm had to be reassigned as part of this manual process. Consider a small example of two physicians d_1 and d_2 and two billets b_1 (California) and b_2 (Japan). Suppose the pilot project assigned d_1 to b_1 and left b_2 and physician d_2 unassigned. Suppose that d_2 cannot be assigned to b_2 since b_2 is overseas and d_2 is currently stationed overseas. The pilot project would require manual reassignment of d_1 to b_2 in order to assign d_2 to b_1. A physician who was successfully matched by the algorithm to a hospital in California could thus be scrambled to a billet in Japan in order to allow for a physician in Japan, unmatched by the algorithm, to come back to CONUS. As of this writing, development of a suitable market design for this reallocation is an open research question.

3.3. Post-Pilot Survey

This survey was approved by the Naval Medical Center San Diego Institutional Review Board. For several months after the match, two of the authors surveyed participants by email or phone. The survey was completed for the last three years of the match. Officers were asked whether they preferred to keep using the deferred acceptance algorithm or to go back to the status quo assignment method. They were also asked their satisfaction with the use of deferred acceptance on a five point Likert scale.

Demographic predictor data, obtained from Navy databases included seniority, gender, and residency graduation status. The type of command a physician was coming from, or matched to, was grouped by two variables: geographic location (CONUS or OCONUS) and position type (shore or ship). The pilot project did not include any administrative billets, and as such the post-pilot survey data includes responses across four billet types: CONUS shore, CONUS ship, OCONUS shore, and OCONUS ship.

We used physicians' rank lists to generate three predictors of wanting to keep the match: number of commands listed, rank given, and rank-pick-match. We hypothesized that the more effort a physician put into the match process, the more likely they would appreciate the match; number of commands listed was used as a surrogate for effort. Rank-given (mean) to a physician was used as a surrogate for the physician's competitiveness and used to test the hypothesis that more competitive physicians would want to keep the match. For example, if one command ranked a physician as that command's first choice, a second command ranked the same physician as its fifth choice, and no other command included that physician on the command's rank list, the physician's rank-given was $3 = (1+5)/2$. Lastly, the *rank-pick-match* (derived from rank of the pick matched to) was used to test the hypothesis that those physicians who matched to a high pick would want to keep it. For example, if a physician matched to a command they ranked as their second choice, their rank-pick-match would be 2. We also report the mean of this number by year as a way to measure overall match success by year. For example, in 2017, the mean rank-pick-match was 2.8. This number indicates that in 2017, on average, physicians gave the command they were assigned a rank of 2.8. A value of 1.0 would indicate that every physician received their first choice. However, the mean rank-pick-match should be used with some caution since it perforce excludes any physician scrambled to a command not on the physician's rank list. In the post-pilot survey, physicians who were scrambled outside of the match to commands not on their rank list generally had a negative response to the pilot project. Thus we hypothesized that those scrambled would not want to keep the match.

Each command appointed a command match representative, usually the department head, who oversaw the command's participation in the pilot project and submitted the command's rank list. Study personnel surveyed the command match representative 1–3 months after the match. In addition to asking them about satisfaction and if they wanted to keep the match, they were also asked if the benefits of the match were worth the extra work, if they had made efforts to recruit physicians, and if they gave preference to physicians coming from ship billets or OCONUS commands. (Commands were instructed that if two candidates were generally equal, to give preference to members coming from ship or OCONUS commands.)

Popularity of command categories were defined by two measures: the number of applicants for each position, and the mean rank given to the command on rank lists. We say a command had a *successful algorithm match* if the algorithm filled all of that command's available billets prior to any scrambling.

4. DATA ANALYSIS

We used R to complete all data analysis. An original repository of data and code of the pre-pilot survey is available at https://tinyurl.com/2euht2p6. We provide a repository of code and data at https://tinyurl.com/685s2cwa for all of the data analysis of the pre-pilot survey, pilot project, and post-pilot survey.

4.1. Pre-Pilot Survey

In this section, we present summary statistics of the survey data, followed by two analyses. First, we estimated the relationship between satisfaction with the status quo assignment process and the seven predictors described in Section 3.1. Second, we studied the mismatch between supply and demand for billets in each of nine geographic regions. We studied this mismatch in two ways. First, we imputed excess demand for each of the regions using survey respondents' reported first geographic choices. Second, we constructed upper and lower demand envelopes for billets by geographic region based on reported preferences and a hypothetical implementation of deferred acceptance. These envelopes show that deferred acceptance may terminate as soon as the third round and thus substantially improve at least geographic outcomes relative to the status quo assignment process.

The response rate was 26.5% (1,015 respondents out of 3,825 Navy Medical Corps officers). Partial completion of the survey resulted in 951 usable observations. Non-surgical and family physicians were over-represented. Operational, resident, and junior (O3) physicians, and physicians in CONUS shore or administrative billets were underrepresented.

Table 1 describes the characteristics of survey respondents and specialty-weighted mean satisfaction by subgroups. On a five point Likert scale, with 5 being the most satisfied, the mean specialty-weighted satisfaction was 3.4. Granular data was not available to weight for officer rate and billet category. The pilot project itself involved emergency physicians; in this group, mean satisfaction was 3.8. These summary statistics show that mean physician satisfaction with the status quo assignment process positively correlates with officer rate, current assignment to a shore or CONUS billet, and certain specialty categories.

	% of Respondents	% of USN Medical Corps	Weighted Mean Satisfaction (std. error)
Physician Specialty	(n = 985)		
Non-surgical	49.5	28.3	3.9 (0.05)
Surgical	17.4	18.3	3.7 (0.09)
Resident	5.4	19.9	3.5 (0.17)
Family Medicine	18.6	9.1	3.4 (0.10)
Operational	9.1	24.4	2.7 (0.14)
Officer Rate	(n = 990)		
O6 (CAPT)	12.8	11.0	4.0 (0.15)
O5 (CDR)	23.4	18.2	4.0 (0.08)
O4 (LCDR)	36.5	32.0	3.5 (0.09)
O3 (LT)	27.3	38.8	2.9 (0.10)
Billet Category	(n = 957)		
CONUS Shore	72.6	78.9	3.7 (0.06)
OCONUS Shore	10.7	8.7	3.5 (0.13)
CONUS Ship	12.4	10.6	2.9 (0.15)
OCONUS Ship	4.3	1.9	2.2 (0.23)
All Respondents			3.4 (0.06)

Notes: USN Medical Corps consists of 3,825 physicians. We weighted mean satisfaction by the distribution of USN Medical Corps physicians by specialty. The unweighted overall mean satisfaction was 3.6.

Table 1: Pre-Pilot Survey: Summary Statistics and Mean Satisfaction with the Status Quo Assignment Process

We asked several other questions about the detailing process in order to understand attitudes towards personnel assignment in general and the status quo assignment process in particular. The results appear in Table 2. A minority of respondents (38.2%) reported that they would prefer to receive their next billet via deferred acceptance. Most USN physicians said that doctors coming from generally less desirable billets or more senior doctors should receive some preference during billet assignment. Curiously, the officers who said that more senior doctors should receive preference in billet assignment had the same average officer rate (4.2) as those respondents who said more senior officers should not receive preference in the assignment process.

One item not directly related to the status quo assignment process itself but of practical concern to USN physicians is the lead time they receive with new orders. Assuming a July move to a new billet, 32.5% of respondents preferred a hypothetical assignment match day in January; December was the second most popular month at 28.7%.

Question	% Yes	% No	n
Would you prefer assignment via deferred acceptance?	38.2	61.8	803
Should doctors coming from OCONUS/ship billets receive preference?	82.3	17.7	820
Should senior doctors receive preference?	75.2	24.8	812

Table 2: Pre-Pilot Survey: Other Attitudes Towards The Assignment Process

Table 3 estimates linear regressions in which we treat the five point Likert outcome as a continuous variable. The five point Likert outcome is a categorical variable, which suggests ordinal logistic regression may be a better choice than linear regression. We therefore completed ordinal logistic regression as a robustness check. Ordinal logistic regression shows that respondents perceived the distances between satisfaction levels 1 and 2; 2 and 3; 4 and 4; and 4 and 5 as roughly equivalent. Since the regressions support the same conclusions and the linear regressions are somewhat easier to interpret, we present and discuss the linear regressions here; the corresponding ordinal logistic regressions appear in Appendix C. Column 2 of Table 3 shows the regression

$$\begin{aligned}\text{Sat} = \beta_0 + \beta_1\text{OfficerRate} + \beta_2\text{SpecialtyType} + \beta_3\text{StaffingRate} + \beta_4\text{BonusPay}\\ + \beta_5\text{LowDeployment} + \beta_6\text{PrefCurrentOrders} + \beta_7\text{BilletType} + \varepsilon\end{aligned}$$

Staffing rate, bonus pay, and low deployment rate are all specific to each specialty; a joint F-test ($p = 0.244$) shows that we cannot reject the hypothesis that $\beta_3 = \beta_4 = \beta_5 = 0$. A subsequent joint F-test ($p = 0.355$) shows that we cannot reject the hypothesis that in addition, $\beta_7 = 0$ for administrative, OCONUS shore and CONUS ship billets. Columns 3 and 4 show these regressions respectively.

These regressions refine the observations from the summary statistics in Table 1. Satisfaction among physicians in surgical, family, and operational

Regressor	Estimate (std. error)	Estimate (std. error)	Estimate (std. error)
Officer Rate			
O5 (CDR)	-0.325 (0.132)**	-0.330 (0.132)**	-0.306 (0.129)**
O4 (LCDR)	-0.403 (0.129)***	-0.424 (0.128)***	-0.387 (0.123)***
O3 (LT)	-0.659 (0.146)***	-0.660 (0.143)***	-0.621 (0.137)***
Specialty Type			
Surgical	-0.138 (0.123)	-0.189 (0.104)*	-0.175 (0.103)*
Family	-0.463 (0.121)***	-0.416 (0.105)***	-0.412 (0.104)***
Operational	-0.654 (0.225)***	-0.480 (0.175)***	-0.551 (0.151)***
Resident	-0.671 (0.269)**	-0.489 (0.199)**	-0.498 (0.198)**
Staffing Rate	0.101 (0.250)		
Bonus Pay	-0.006 (0.008)		
Low Deployment Rate	0.189 (0.111)*		
Preference for Current Orders	-0.530 (0.034)***	-0.530 (0.035)***	-0.529 (0.033)***
Current Billet Type			
Administrative	-0.148 (0.176)	-0.150 (0.174)	
OCONUS Shore	0.132 (0.128)	0.137 (0.128)	
CONUS Ship	-0.056 (0.141)	-0.080 (0.138)	
OCONUS Ship	-0.613 (0.219)**	-0.628 (0.218)***	-0.606 (0.212)***
Observations	751	751	751

Notes: 200 observations dropped due to missing values. Preference for Current Orders is an integer in $\{1, 2, 3, 4, 5\}$ corresponding to the respondent's description of his/her/their current assignment as his/her/their 1st, 2nd, 3rd, 4th, or 5th or lower choice. The reference categories for Officer Rate, Specialty Type, and Deployment Rate were O6 (CAPT), Nonsurgical, and High respectively. In columns 2 and 3, the reference category for Current Billet Type was CONUS Shore. In column 4, we use a single dummy for OCONUS Ship billets; the reference category is non-OCONUS Ship billets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 3: Pre-Pilot Survey

specialties, as well as among residents is substantially lower than satisfaction among physicians in non-surgical specialties. Unsurprisingly, respondents who receive more preferred orders are more satisfied with the status quo assignment process than respondents who receive less preferred orders. Physicians serving in OCONUS ship billets are also substantially less satisfied with the status quo assignment process.

We elicited geographic preferences from survey respondents. USN physician billets are distributed across nine regions. Table 4 illustrates the difference between each region's geographic desirability and the availability of positions in those regions. Based on 916 responses to Question D3 in the pre-pilot survey, we imputed excess demand in each region in Table 4 by

$$\text{Imputed Excess Demand} = (\#\text{Physicians Ranking Region 1st})\frac{4009}{916} - \#\text{Billets in Region}$$

Region	# Billets	# Physicians Ranking 1st	Imputed Excess Demand
Europe	120	136	475
Pacific NW	200	75	128
Northeast	40	27	78
Southern California	1203	283	36
Middle East / Africa	40	5	−18
Southeast (NC, FL)	481	102	−35
Pacific (HI, Guam, Japan)	441	77	−104
Washington DC Area	682	115	−179
Southeast VA	802	96	−382
Total	4009	916	0

Table 4: Imputed Excess Demand

Negative excess demand may naturally be interpreted as excess supply. Europe had the smallest share of available positions (2.9%) compared to the share of respondents who listed Europe as their first choice (14.8%). Billets in Southern California were the most common first choice, with 30.9% of respondents ranking Southern California as their first geographic choice; 30.0% of physician billets are in Southern California, which implies a modest excess demand of 36. Billets in Southeast Virginia were the least popular: 10.5% of respondents ranked this area as their top choice, while 20.0% of the available billets are in this region.

Table 5 further characterizes the supply and demand relationship via a series of demand envelopes which we constructed using hypothetical billet assignment via physician-proposing deferred acceptance. Unfortunately, due to data limitations we could not extend this analysis to include other billet characteristics. We first created a scaled supply of billets in each region according to

$$\text{Supply of Billets in Region} = (\#\text{Billets in Region})\frac{916}{4009}$$

Second, we counted the smallest and largest number of possible proposals to each region in three rounds of deferred acceptance; we stopped at the third round since it is the earliest round in which the supply of billets in each region is between the minimum and maximum demand, and thus the earliest round in which deferred acceptance with respect to location alone can terminate. A more detailed description of our methodology appears Appendix B. The demand envelopes show that the Washington DC Area and Southeast Virginia are the least popular regions: they are the only two regions with a excess supply of billets at the end of the second round of deferred acceptance.

		Demand (# of Proposals in Deferred Acceptance)				
Region	Supply	Round 1	Round 2		Round 3	
			Min	Max	Min	Max
Europe	27	136	27	50	27	219
Pacific NW	46	75	46	68	46	257
Northeast	9	27	9	33	9	72
Southern California	275	283	288	348	275	423
Middle East / Africa	9	5	5	11	5	27
Southeast (NC, FL)	110	102	102	135	102	212
Pacific (HI, Guam, Japan)	101	77	93	132	93	262
Washington DC Area	156	115	115	139	115	256
Southeast VA	183	96	96	124	96	235

Table 5: Demand Envelopes from Deferred Acceptance

4.2. The Pilot Project and Post-Pilot Survey

Table 6 provides demographic data of officers participating in the pilot project. Overall, 24.7% of participants were female, 39.4% were new residency graduates, and the majority of participants (79.1%) were junior officers (rate O3 or O4).

	2017	2018	2019	2020	Overall
Participants	46	58	59	68	231
% Female	28.3	22.4	16.9	30.9	24.7
% Rank					
O3	45.7	25.9	28.8	28.8	33.0
O4	34.8	46.6	52.5	52.5	46.1
O5	15.2	19.0	11.9	11.9	14.8
O6	4.3	8.6	6.8	6.8	6.1
% New Residency Graduate	52.2	34.5	44.1	30.9	39.4
Mean # of commands ranked (std. deviation)	8.6 (4.7)	5.7 (3.8)	7.0 (5.5)	6.2 (4.1)	6.7 (4.6)

Table 6: Pilot Project: Distribution of Participants

Table 7 illustrates various officer outcomes by year. In particular, the first row shows that most physicians prefer deferred acceptance (the assignment method of the pilot project modulo the scramble) to the status quo assignment process (via detailers). The algorithm successfully matched 81.8% of participants; however, 31.2% of participants were scrambled. Only 10.9% of participants were scrambled into a position that was not listed on their rank list. The *mean rank-pick-match* is the rank of the pick the physician was assigned to; on average, physicians were assigned to their second ranked command. Fully 78.7% of respondents want to keep using the match for assignments.

We use logistic regression to study which physicians preferred the pilot project to the status quo assignment process. Potential regressors included gender, new graduation status, seniority, matched to command category (CONUS ship, OCONUS shore, OCONUS ship, and CONUS shore), current command category, number of commands listed, rank given, rank-pick-match, successful match, and scrambled status (yes or no, regardless of whether the physician was scrambled to a command on his/her/their rank list).

In practice, rank-pick-match is a poor regressor: since some physicians were scrambled off of their rank lists, rank-pick-match is not well-defined for those physicians. However, we can use rank-pick-match to construct a well-defined regressor for all physicians: if rank-pick-match is not well-defined we say the physician was scrambled off list; otherwise we say the physician was not scrambled off list (either because the physician was not scrambled at all, or because the physician was scrambled to a command on his/her/their rank list. These three variables (successful match, scrambled status, rank-pick-match)

	2017	2018	2019	2020	Overall
% prefer pilot project	No survey this year	78.6 RR=96.6	73.9 RR=78.0	88.0 RR=36.8	78.7 RR=55.0
% successful algorithm match	82.6	81.0	78.0	85.3	81.8
% scrambled	28.3	34.5	37.3	25.0	31.2
% scrambled off rank list	2.2	12.1	15.3	11.9	10.9
Mean rank-pick-match	2.8 (2.0-3.6)	2.2 (1.6-2.8)	2.2 (1.6-2.7)	2.1 (1.5-2.6)	2.3 (2.0-2.6)
Satisfaction	No survey this year	3.9 (3.6 - 4.3) RR=98.3	3.7 (3.3-4.1) RR=84.7	3.9 (3.4 - 4.4) RR=50.0	3.8 (3.6-4.1) RR=76.2

Notes: RR = Response Rate. Satisfaction was measured on a five point Likert scale in which 1 = not satisfied and 5 = very satisfied. We report 95% confidence intervals in parentheses for continuous variables.

Table 7: Officer-side outcomes by year

collectively identify one of five possible pilot project outcomes for a physician; we therefore consolidate them into a single categorical variable *pilot outcome* that takes one of the following five values:

1. YNN (Successful Match = “yes”, Scrambled = “no”, scrambled off of rank list = “no”) Successfully matched to a command (necessarily on his/her/their rank list) by the pilot project and not scrambled elsewhere
2. YYN (Successful Match = “yes”, Scrambled = “yes”, scrambled off of rank list = “no”) Successfully matched to a command (necessarily on his/her/their rank list) by the pilot project and scrambled to a different command on his/her/their rank list
3. YYY (Successful Match = “yes”, Scrambled = “yes”, scrambled off of rank list = “yes”) Successfully matched to a command (necessarily on his/her/their rank list) by the pilot project and scrambled to a command not on his/her/their rank list
4. NYN (Successful Match = “no”, Scrambled = “yes”, scrambled off of rank list = “no”) Not matched to a command by the pilot project and scrambled to a command on his/her/their rank list
5. NYY (Successful Match = “no”, Scrambled = “yes”, scrambled off of rank list = “yes”) Not matched to a command by the pilot project and scrambled to a command not on his/her/their rank list

Table 8 shows the univariate odds ratios for each of successful match, scrambled status, and rank-pick-match, as well as the constructed variable pilot outcome. Based on these regressions, the discussion above, and subsequent robustness checks, we use pilot outcome in all logistic regression analyses in the main body of the paper.

Regressor	Estimate (std. error)	Estimate (std. error)	Estimate (std. error)	Estimate (std. error)
Successful Match				
Yes	3.346 (1.690)**			
Scramble				
Yes		0.267 (0.120)***		
Scrambled Off List				
Yes			0.125 (0.078)***	
Pilot Outcome				
NYN				0.267 (0.161)**
YYN				1.600 (1.750)
NYY				0.160 (0.140)**
YYY				0.064 (0.057)***
Observations	127	127	127	127

Notes: 104 observations dropped due to missing values. The reference category for Pilot Outcome was YNN, i.e., physicians who were successfully matched to a command (necessarily on his/her/their rank list) by the pilot project and not scrambled elsewhere. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: Post-Pilot Survey: Pilot Project Outcomes

Column 2 of Table 9 shows our estimate of the following logistic regression model:

$$
\begin{aligned}
\frac{\Pr(\text{Prefer})}{1-\Pr(\text{Prefer})} = \exp[&\beta_0 + \beta_1\text{Gender} + b_2\text{NewGrad} + \beta_3\text{OfficerRate} \\
&+ \beta_4\text{CurrentCommand} + \beta_5\text{MatchedTo} \\
&+ \beta_6\text{NumberOfCommandsListed} + \beta_7\text{MeanRankByCommands} \\
&+ \beta_8\text{PilotOutcome}]
\end{aligned}
$$

The full regression is somewhat difficult to interpret due to the small size of the dataset. However, physicians for whom the pilot project did not work, i.e., PilotOutcome was YYY typically did not prefer the pilot project. We conjecture that these physicians felt that the pilot project did not live up to its promise of making assignments via deferred acceptance, and moreover did not respect their reported preferences since they were scrambled off of their rank lists. A similar conjecture may hold for physicians whose PilotOutcome was NYN, i.e., who were not successfully assigned to a command via deferred acceptance even if the scramble subsequently assigned them to a command on their rank lists.

Regressor	Estimate (std. error)	Estimate (std. error)	Estimate (std. error)
Male	0.499 (0.388)		
New Graduate (Yes)	2.323 (2.424)		4.546 (3.194)**
Officer Rate			
O5 (CDR)	0.175 (0.262)		
O4 (LCDR)	1.504 (2.239)		
O3 (LT)	0.835 (1.346)		
Current Command			
CONUS Ship	0.673 (0.674)		
OCONUS Shore	0.460 (0.402)		
OCONUS Ship	0.091 (0.113)*		
Matched To			
CONUS Ship	2.013 (1.976)	1.559 (1.188)	1.371 (1.084)
OCONUS Shore	0.290 (0.203)*	0.418 (0.251)	0.317 (0.194)*
OCONUS Ship	0.109 (0.155)	0.284 (0.330)	0.091 (0.119)*
Number of Commands Listed	0.919 (0.074)	1.010 (0.066)	
Mean Rank by Commands	0.928 (0.061)	0.952 (0.054)	
Pilot Outcome			
NYN	0.104 (0.087)***	0.202 (0.139)**	0.208 (0.144)**
YYN	1.835 (2.401)	2.265 (2.584)	1.806 (2.072)
NYY	0.202 (0.225)	0.181 (0.188)*	0.161 (0.165)*
YYY	0.052 (0.069)**	0.113 (0.122)**	0.108 (0.127)*
Observations	123	123	123

Notes: 108 observations dropped due to missing values. The reference category for Pilot Outcome was YNN, i.e., physicians who were successfully matched to a command (necessarily on his/her/their rank list) by the pilot project and not scrambled elsewhere. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Post-Pilot Survey: Logistic Regressions

Column 3 shows the logistic regression restricted to regressors arising directly from the pilot project, i.e., assuming the coefficients of the demographic regressors are $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. The estimated odds ratios remain similar to the corresponding coefficients in Column 2. Finally, in Column 4 we show the logistic regression using backwards model selection. Backwards selection retains NewGrad as the only demographic regressor and removes both NumberOfCommandsListed and MeanRankByCommands. We hypothesize that the large and significant odds ratio for NewGrad is due to (at least) a perception that the pilot project provides new USN physicians with access to a larger number of billets that they find desirable. It is possible that NumberOfCommandsListed and MeanRankByCommands are not significant since they are not pilot project outcomes per se. The pre-pilot survey showed that residents and junior officers are relatively dissatisfied with the status quo assignment process, which is consistent with a reported preference for continuing to determine billet assignments via the pilot project methodology. As noted elsewhere, the small sample size suggests caution in interpreting the odd ratios, and that a larger future study would shed more light on which USN physicians truly prefer assignment via some sort of market design relative to the status quo assignment process.

Otherwise identical regressions to those in Table 9 using the individual categorical variables (successful match, scrambled, and scrambled off list) instead of pilot outcome appear in Table 14 in Appendix D as robustness checks. The odds ratios of the other regressors, i.e., neither PilotOutcome nor its component categorical variables, are similar in the two regressions. Due to the small number of observations, we also estimated all of the univariate logistic regressions, which appear in Table 8 or else in Tables 15 and 16 in Appendix D. The univariate odds ratios for MatchedTo highlight the fact (also evident to some extent in the larger regression of Table 9) that assignment to an OCONUS command was a reasonable indicator of a wish to stick with the status quo assignment method.

Table 10 shows the command-side outcomes of the pilot project, grouped by command category. CONUS shore commands generally fared better than other commands, and OCONUS ship commands generally fared worse. CONUS shore commands were significantly more likely than OCONUS ship commands to have a successful algorithm match, received significantly more applicants per position, and were consistently ranked higher by USN physicians. It is theoretically possible that the least talented individuals would fill the less popular billets, but that outcome did not occur: our surrogate marker for talent was the mean rank of selected individuals and there was no difference in this marker between command categories.

	CONUS Shore	CONUS Ship	OCONUS Shore	OCONUS Ship	Overall	Overall *p*-value
% Available positions	46.3	22.1	23.8	7.8	100.0	N/A
% Successful algorithm match	92.9	78.9	48.1	28.6	69.1	<0.001
Applicants per position	10.8 (7.7–13.8)	7.9 (6.3–9.5)	7.5 (5.7–9.3)	4.1 (1.5–6.7)	8.4 (7.1–9.7)	0.001
Mean rank given to command	4.7 (4.2–5.1)	6.0 (5.1–7.0)	7.3 (6.7–7.9)	10.0 (7.4–12.6)	6.3 (5.8–6.8)	<0.001
Mean rank of selected individuals	7.3 (6.5–8.0)	7.9 (6.7–9.1)	7.9 (6.6–9.1)	7.1 (5.5–8.7)	7.5 (7.0–8.0)	0.707

Notes: There were 231 available billets. We report 95% confidence intervals in parentheses. We use Bonferroni correction for all individual pairwise comparisons.

The share of CONUS shore commands with a successful algorithm match was significantly greater than share of OCONUS shore and OCONUS ship commands with a successful algorithm match.

We analyzed overall and pairwise comparisons on ln(applicants per position). CONUS shore and CONUS ship commands received significantly more applicants per postion than OCONUS ship commands.

We analyzed overall and pairwise comparisons on ln(mean rank). CONUS shore commands were significantly more popular (lower rank given) than CONUS ship commands, which were significantly more popular than OCONUS shore commands, which were significantly more popular than OCONUS ship commands.

Table 10: Command-side outcomes by command category

Table 11 shows the command-side survey results, grouped by command category. A successful algorithm match, a desire to keep using the match, satisfaction, and command popularity (measured by a applicants/position and mean rank given to command) all showed a high to low gradient moving from CONUS shore, through CONUS ship and OCONUS shore, ending at OCONUS ship commands. The match was less successful and less popular at ship and OCONUS commands. All command categories generally thought the benefits of the match were worth the extra work required. Only 34.5% of commands took active steps to recruit officers, and 81.1% of commands gave preference to officers coming from ship or OCONUS commands.

	CONUS Shore	CONUS Ship	OCONUS Shore	OCONUS Ship	Overall	Overall p-value
% prefer pilot project	90.0 RR=90.9	100.0 RR=85.7	75.0 RR=76.2	100.0 RR=20.0	87.8 RR=79.0	0.009
Satisfaction	4.4 (4.0–4.8) RR=91.0	4.3 (3.7–4.8) RR=85.7	3.3 (2.6–3.9) RR=76.2	2.3 (-0.5–5.2) RR= 60.0	3.9 (3.6–4.2) RR=82.3	0.001
% Pilot project was worth the extra work	90.0 RR=90.9	100.0 RR=85.7	75.0 RR=76.2	100.0 RR=40.0	88.0 RR=80.6	0.070
% Recruited	33.3 RR=40.9	37.5 RR=57.1	27.3 RR=52.4	100.0 RR=20.0	34.5 RR=46.8	0.710
Gave preference for OCONUS or ship	76.9 RR=59.1	90.0 RR=71.4	76.9 RR=61.9	100.0 RR=20	81.1 RR=59.7	0.529

Notes: We report 95% confidence intervals in parentheses and response rates. We use Bonferroni correction for all individual pairwise comparisons.

The percentage of OCONUS ship commands that preferred the pilot project to the status quo assignment process was significantly greater than the percentage of CONUS shore commands exhibiting the same preference.

We analyzed overall and pairwise comparisons on ln(satisfaction). CONUS shore commands were significantly more satisfied than OCONUS shore and OCONUS ship commands. CONUS ship commands were significantly more satsified than OCONUS ship commands.

Table 11: Command-side post-pilot survey responses by command category

We use the Pearson χ^2 test to compare means of binary variables across command categories to detect an overall p-value; if less than 0.05, we use Fisher's Exact Test with Bonferroni correction to complete individual pairwise comparisons. We use 1-Way ANOVA to compare means of continuous variables between command categories to detect an overall p-value; if less than 0.05, we use pairwise t-tests with Bonferroni correction to complete individual pairwise comparisons.

Finally, we offer a comparison of the pilot project outcomes to NRMP outcomes. NRMP is a longstanding and reasonably successful matching market. NRMP publishes annually the match rate by choice and the unmatched rate for American medical school seniors and independent applicants. Independent applicants, generally considered less competitive, include American medical school graduates (usually applying for the second time) and international graduates. We compared these results to our pilot project. Table 12 illustrates the unmatched rate, and the percent of matches by choice, of our pilot project compared to NRMP results. Our match performed similarly to the rate of physicians in the last year of medical school, and better than independent applicants.

	Pilot Project (2017–2020)	NRMP (MD Seniors)	NRMP (Independent Applicants)
1st choice	53.0	46.3	29.0
2nd choice	11.7	15.4	11.9
3rd choice	7.4	9.4	7.2
4th choice	6.5	6.4	4.7
> 4th choice	10.4	16.5	8.7
Unmatched	10.9	5.9	38.4

Table 12: Match Rate Comparison: Pilot Project vs. NRMP

4.3. Limitations

Our pre-pilot survey respondents were not perfectly representative of our target population. Younger physicians were underrepresented. It was difficult to get responses from the O3 officers who make up a large proportion of operational and resident physicians. We distributed our survey using leaders within each specialty. This was an effective method for specialty trained physicians who are usually ranked O4 and above. The O3 physicians are usually in operational units which are not tightly associated with the specialty, or they are in the midst of their training and primarily focused on their education. We report the results of a weighted analysis which reduced the overall mean satisfaction slightly.

There are inherent limitations in the use of military personnel databases. Different offices use different databases and variable coding, which can lead to disparate information. Our data had face validity, and we think it unlikely that database or coding variability external to our study affected our results.

As a cross-sectional study, this represents attitudes at one point in time, and our results may not be generalizable to other years. The year this survey was completed, a budget shortfall delayed the release of orders; this caused significant anger among personnel, because without official orders, the moving process could not be initiated and executed. This may have accentuated the selection bias frequently seen in surveys and led to lower satisfaction scores.

This survey of USN physicians may have limited generalizability; however, it has relevance to the military at large. First, changing assignments as a career progresses is common in most militaries. As noted earlier, Korkmaz et al. (2008) describe a theoretical model utilizing matching principles designed to improve military assignment policy. Second, this survey showed that remote (OCONUS) locations are less desirable. This is likely because families prefer to settle down, keep their children in the same schools, buy a house, and spouses often have jobs that benefit from stability. Moving anywhere makes this a challenge, but it is particularly challenging when living in remote or international areas. Thus, these issues would not be unique to USN physicians.

The pilot project itself and subsequent analysis had no control group. While it did establish baselines for comparison, there was no direct comparison with the standard assignment system. It would be possible to do a cluster randomized trial in the military setting; different communities (pilots, surface warfare officers, etc.) could be the unit of randomization. Ideally, this would be pursued.

The post-pilot survey response rate was good in 2018 and 2019 but lagged in 2020. Our study was disrupted by the COVID pandemic. If 2020 was a particularly good or bad year, that could bias results in either direction. The market rules and participant samples were similar across years, so we think it unlikely that significant error was introduced.

There was likely selection bias in the survey response, but it is not clear how much it affected the mean and variance of the responses. One could imagine that individuals with either strongly negative or strongly positive feelings would be more likely to respond than neutral individuals.

The pre-pilot survey was anonymous but the post-pilot survey was completed by the study investigators by email and phone; respondents were not anonymous. As the study investigators oversaw the respondents' assignments, we were in a position of power. This method could have biased our results in favor of the match. However, respondents' were aware the investigators were not in a position to affect any aspect of the respondents' professional outcomes and therefore had little to no incentive to distort their responses.

Lastly, the sample size for our intervention was small at 231 billets. This project was unfunded and completed in a skeptical environment; thus, it was limited to the emergency medicine community over four years. The project was abandoned when one of the coauthors rotated to a new position and was no longer able to champion the project. Respondents' preference to continue using deferred acceptance for personnel assignment points to misplaced skepticism.

5. CONCLUSION

This is the first (to the authors' knowledge) study in the peer-reviewed literature describing participant satisfaction in a matching market for USN physicians. The post-pilot survey showed that approximately 78.7% of officers, and 87.8% of commands wanted to keep using the deferred acceptance algorithm as part of USN physician assignment process, despite initial skepticism in the pre-pilot survey. The outcome of the pilot project compared favorably with NRMP. Predictably, those who matched to desirable commands were happy with the match; also, popular commands want to keep using the match.

Military leadership will want to know, "Did the pilot project work?" It is difficult to give a definitive outcome to answer this question. Our primary outcome, asking participants if they liked it, seems reasonable. By this measure, despite some initial skepticism demonstrated in the pre-pilot survey, the match seemed popular with both officers and commands. NRMP uses the percent of matches by choice (i.e., how many get their first choice) as a performance metric. Our results compare favorably to NRMP by this measure.

In addition to a measured outcome, military leaders can draw on the large body of literature in other matching markets to gain insight, and likely confidence, that the principles of matching markets can be applied to the military personnel system. However, design will need to incorporate the constraints inherent in the military system.

We hypothesized that there would be talent concentration in the most popular billets. However, there was no difference in the mean rank of selected individuals between command categories. This is inconsistent with talent concentration but does not rule it out. Competitive individuals may have known they would match at popular commands and only listed these; thus, those less competitive individuals may have a mean rank biased upward since they are not always compared to competitive individuals at all sites. Considering the disparity of command popularity, we suspect some degree of talent concentration.

This work strongly illustrates that OCONUS and ship billets are the least popular. The pre-pilot survey revealed that USN physicians serving overseas or in ship billets are significantly less satisfied with their assignments than their peers. During the pilot project, OCONUS ship billets were, by far, the least popular; also, officers coming from or going to these billets were the least satisfied with the match. This result has face validity. In our anecdotal experience as detailers, we consistently found that the least desirable billets are OCONUS and ship. Active duty physicians, in general, like to be stationed in CONUS, and like to work in traditional medical settings like hospitals and clinics.

This is a challenge for military leaders because OCONUS and ship billets are a high priority; stateside clinic jobs can be filled with civilians. As long as there are less-desirable jobs that need to be filled, the military personnel assignment system will, to some extent, rely on coercion. A coercive environment is unlikely to attract top talent into military service. Also, decreased job satisfaction likely contributes to physician burn-out, and burn-out is associated with poorer-quality healthcare and reduced safety for patients. (Salyers et al., 2017) Ideally, the highest priority billets would be filled by officers who want to be there.

Navy leadership should investigate why these positions are less desirable. Anecdotally, we know that many physicians serving overseas are concerned about skill retention. OCONUS populations are relatively healthy due to the overseas screening process for both military members and families, and patient volumes are low. Also, it can also be difficult for the professional lives of military spouses to move overseas. Measures specific to these concerns, such as shortened tour lengths (which are already used to some extent) might make these positions more desirable. While this work just focuses on military physicians, we suspect, based on anecdotal experience, this is generalizable to other military communities.

Instead of forcing physicians into assignments, it might be possible, through incentives, to make ship and OCONUS positions more desirable. Financial incentives are theoretically simple. Additionally, promotion boards could be directed to promote those physicians who have served OCONUS or in ship billets. Lastly, personnel coming from less-desirable positions could have priority for more desirable positions at their next opportunity for assignment. This last idea was supported by 82.3% of respondents in the pre-pilot study.

In the NRMP scramble that occurs after algorithm matches are announced, unmatched physicians and unfilled residency positions seek each other after the fact. The scramble may be unpleasant for those involved, but its existence does not interfere with the NRMP algorithm's match: it is acceptable to NRMP if some physicians do not match, or if some positions remain unfilled. The scramble in our match is different: in the military, participants cannot be left unmatched as they are under a contractual obligation to serve and the military is under obligation to employ its members. Additionally in the context of USN physicians, USN cannot leave high-priority OCONUS ship billets unfilled. This constraint led to our military match scramble, a feature of our match viewed unfavorably. During the survey process, many remarked that the positive impression of a transparent match was undone by the unsystematic scramble process. While we think some degree of scramble is inevitable, the goal should be to minimize it and make the process transparent. Future work, both theoretical and practical, should focus on a scramble that is as fair as possible while still meeting the needs of USN.

A. SURVEY TEXT

A.1. Pre-Pilot Survey

1. What is your primary specialty?
2. What is your rank?
3. What year did you execute orders for your current billet? (Consider retour orders the same as a PCS set of orders.)
4. How would you describe your current billet?
5. For your last set of orders, how many months prior to your move were your orders released? That is, how many months did you have to prepare for your PCS?
6. On a scale of 1 to 5, with 1 being not transparent and 5 being very transparent, how would you rate the transparency of your detailing experience for your last set of orders?
7. On a scale of 1 to 5, with 1 being unsatisfied and 5 being very satisfied, how would you rate your overall detailing experience for your last set of orders?

8. On a scale of 1 to 5, with 1 representing a significant problem and 5 being not a problem at all, how would you rank the problem of favoritism in the billet assignment process?
9. Describe your current assignment:
10. Please rank your desired billet locations with the top level being the most desirable, and the bottom being the least desirable. [Ranking 1, 2, ..., 9]
11. Which career path do you want to pursue in the next 5–10 years?
12. Neither the Army nor the Air Force have physicians in the detailer role. Instead, they have nurses or medical administrators work with specialty leaders to determine assignments. This is different from the current Navy Medical Corps billet assignment process where the detailer is a physician. Would you approve if the detailer position was filled by a non-physician?
13. How long should an individual be allowed to remain at one command?
14. Do you think that there is a problem in the Medical Corps with members not moving? That is, are there too many physicians who get to stay in one place too long?
15. Civilian medical residency positions are assigned using the National Residency Match Program where members submit a preference list, residency directors submit a preference list, and a computer algorithm optimizes a match. This is different from the current Navy Medical Corps billet assignment process where the detailer and specialty leader take input from medical officers and then make a decision. Of these two options, which would you prefer for your military billet assignment?
16. The later the match day, the more information one has before creating their rank list. The earlier the match day, the sooner one can have certainty and prepare. Assuming your were scheduled to execute new orders in July of 2017, what month would you want the match to occur in?
17. Do you think members who are coming from operational or OCONUS assignments should be given preference in billet assignment?
18. Do you think members with more seniority (as defined by time in service or rank) should be given preference in billet assignment?

19. Any last thoughts or input regarding the billet assignment process?

A.2. Post-Pilot Officer Survey

1. On a scale of 1 to 5, with 1 being nontransparent and 5 very transparent, how would you rate the transparency of the detailing match?
2. On a scale of 1 to 5, with one being not satisfied, and five being very satisfied, how would you rate your overall satisfaction, from a command standpoint, with the detailing match?
3. When considering the qualities of a potential duty station, rank the following three characteristics in terms of importance: geographic location, clinical skills growth, and Navy career growth?
4. If you had to choose between the prior assignment system where the specialty leader and the detailer make an assignment decision while considering the desires of the constituent, or the new match system, what would you prefer?
5. What are the drawbacks to the match?
6. What are the advantages to match?

A.3. Post-Pilot Command Survey

1. On a scale of 1 to 5, with 1 being nontransparent and 5 very transparent, how would you rate the transparency of the detailing match?
2. On a scale of 1 to 5, with one being not satisfied, and five being very satisfied, how would you rate your overall satisfaction, from a command standpoint, with the detailing match?
3. When considering the qualifications of a potential candidate, rank the following three characteristics in terms of importance: Service reputation, skill set, grade.
4. If you had to choose between the prior assignment system where the specialty leader and the detailer make an assignment decision while considering the desires of the constituent, or the new match system, what would you prefer?

5. The match system requires extra work for commands: resumes have to be reviewed, interviews conducted, rank lists comprised. Was this extra work worth it?
6. Were candidates from OCONUS or operational billets given preference? That is, if candidates were roughly equal, were the candidates coming from OCONUS or operational billets given preference?
7. Any suggestions for officers to improve their chances of being selected for your command?
8. Any general thoughts regarding the match?
9. What are the drawbacks to the match?
10. What are the advantages to match?
11. Did you recruit officers you desired at your command, or did you just review the officers who showed an interest in your command?

B. DEMAND ENVELOPE METHODOLOGY

This section describes the construction of the upper demand envelopes presented in Table 5.

First, we scaled the 4,009 billets down to a representative 916 billets according to

$$\text{Scaled \# of Billets in Region} = (\#\text{Billets in Region})\frac{916}{4009}$$

We rounded the scaled number of billets in each region to the nearest integer. Column 2 of Table 5 shows the total number of billets and the scaled number of billets. We refer to the scaled number of billets as the supply in the rest Appendix B.

In each round of physician proposing deferred acceptance, we used the scaled number of billets as the supply in each region. Let R1 (R2, R3) denote round 1 (round 2, round 3) of physician proposing deferred acceptance. Let

1. $|L1|$ denote the number of physicians who listed the location $L1$ as their first choice
2. $|L1.L2|$ denote the number of physicians who listed location $L1$ as their first choice and location $L2$ as their second choice

3. $|L2.L3|$ denote the number of physicians who listed location $L2$ as their second choice and location $L3$ as their third choice

R1: Each physician proposes to his/her/their favorite billet; the number of physicians ranking a region first is the R1 demand in each region. Four regions (Europe, Pacific NW, Northeast, and Southern California) have excess R1 demand while the remaining five have excess supply. Europe, Pacific NW, Northeast, and Southern California reject 109, 29, 18, and 8 proposals respectively, while each of the remaining five regions holds all of the first round proposals it received. All first round proposals tentatively held at the end of R1 are repeated in R2.

R2: For L1 one of Europe, Pacific NW, Northeast, Southern California, and L2 any region not equal to L1, we compute the minimum number of R2 proposals to L2 from physicians ranking L1 first and L2 second as follows:

$$\begin{aligned}\text{min \# proposals to L2 from } |L1.L2| = \min\{0, |L1.L2| \\ -\text{\# of proposals held by L1 in R1}\}\end{aligned} \tag{1}$$

For L1 one of Europe, Pacific NW, Northeast, Southern California, and L2 any region not equal to L1, we compute the maximum number of round 2 proposals to L2 from physicians ranking L1 first and L2 second as follows:

$$\begin{aligned}\text{max \# proposals to L2 from } |L1.L2| = \min\{|L1.L2|, \\ \text{\# of rejections by L1 in R1}\}\end{aligned} \tag{2}$$

In each region, we obtain the minimum (maximum) R2 demand by summing the number of proposals held in R1 together with the minimum (maximum) number of new proposals described in Equations (1) and (2).

Up to seven regions (Europe, Pacific NW, Northeast, Southern California, Middle East / Africa, Southeast (NC, FL), and Pacific (HI, Guam, Japan)) have excess demand, i.e., the maximum number of proposals in R2 exceeds the supply of billets in those regions.

R3: For L2 one of Europe, Pacific NW, Northeast, Southern California, Middle East / Africa, Southeast (NC, FL), and Pacific (HI, Guam, Japan), and L3 any region not equal to L2, we verified that

$$|L2.L3| \leq \text{\# of proposals held by L2 in R2},$$

so that the minimum number of new proposals in R3 to each region is zero.

For L2 one of Europe, Pacific NW, Northeast, Southern California, Middle East / Africa, Southeast (NC, FL), and Pacific (HI, Guam, Japan), and L3 any region not equal to L2, we computed the maximum number of R3 proposals to L3 from physicians ranking L2 second and L3 third as follows:

$$\begin{aligned} &\text{max \# proposals to L3 from } |L2.L3| = \min\{|L2.L3|, \\ &\quad \text{max \# of rejections by L2 in R2}\} \end{aligned} \tag{3}$$

In each region, the minimum R3 demand is the number of proposals held in R2. We obtain the maximum R3 demand by summing the number of proposals held in R2 together with the maximum number of new proposals described in Equation (3).

We terminated the analysis in R3, the earliest round such that supply is between minimum and maximum demand in every region, and therefore the earliest possible round in which physician proposing deferred acceptance can terminate.

C. ROBUSTNESS CHECKS: PRE-PILOT SURVEY

Table 13 shows that ordinal logistic regressions, the points on the five point Likert scale are approximately equidistant, supporting the use of linear regression.

Regressor	Estimate (std. error)	Estimate (std. error)	Estimate (std. error)
Officer Rate			
O5 (CDR)	-0.802 (0.264)***	-0.802 (0.262)***	-0.752 (0.255)***
O4 (LCDR)	-0.966 (0.258)***	-0.981 (0.256)***	-0.910 (0.243)***
O3 (LT)	-1.487 (0.286)***	-1.441 (0.279)***	-1.363 (0.265)***
Specialty Type			
Surgical	-0.133 (0.226)	-0.283 (0.190)	-0.258 (0.189)
Family	-0.784 (0.225)***	-0.736 (0.193)***	-0.720 (0.191)***
Operational	-1.163 (0.414)***	-0.798 (0.327)**	-0.880 (0.279)***
Resident	-1.223 (0.485)**	-0.792 (0.349)**	-0.806 (0.348)**
Staffing Rate	0.380 (0.486)		
Bonus Pay	-0.018 (0.015)		
Low Deployment Rate	0.267 (0.205)		
Preference for Current Orders	-0.913 (0.071)***	-0.909 (0.071)***	-0.904 (0.069)***
Current Billet Type			
Administrative	-0.293 (0.331)	-0.280 (0.328)	
OCONUS Shore	0.215 (0.238)	0.230 (0.238)	
CONUS Ship	-0.063 (0.263)	-0.076 (0.259)	
OCONUS Ship	-1.252 (0.424)***	-1.249 (0.423)***	-1.226 (0.413)***
1 | 2	-5.825 (0.679)***	-6.004 (0.349)***	-5.932 (0.337)***
2 | 3	-4.583 (0.661)***	-4.765 (0.309)***	-4.695 (0.296)***
3 | 4	-3.250 (0.650)***	-3.434 (0.281)***	-3.365 (0.267)***
4 | 5	-1.823 (0.643)***	-2.011 (0.264)***	-1.944 (0.249)***
Observations	751	751	751

Notes: 200 observations dropped due to missing values. Preference for Current Orders is an integer in $\{1, 2, 3, 4, 5\}$ corresponding to the respondent's description of his/her/their current assignment as his/her/their 1st, 2nd, 3rd, 4th, or 5th or lower choice. In columns 2 and 3, the reference categories for Specialty Type, Deployment Rate, and Current Billet Type are Nonsurgical, High, and CONUS Shore respectively. In column 4, we use a single dummy for OCONUS Ship billets; the reference category is non-OCONUS Ship billets. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 13: Pre-Pilot Survey, Ordinal Logistic Regressions

D. ROBUSTNESS CHECKS: POST-PILOT SURVEY

As a robustness check, Table 14 shows regressions identical to those in Table 9, using the individual categorical variables (successful match, scrambled, and scrambled off list) instead of pilot outcome.

Regressor	Estimate (std. error)	Estimate (std. error)	Estimate (std. error)
Male	0.727 (0.520)		
New Graduate (Yes)	3.458 (3.456)		4.728 (3.311)**
Officer Rate			
O5 (CDR)	0.228 (0.318)		
O4 (LCDR)	1.350 (1.860)		
O3 (LT)	0.705 (1.055)		
Current Command			
CONUS Ship	0.896 (0.865)		
OCONUS Shore	0.619 (0.524)		
OCONUS Ship	0.120 (0.145)*		
Matched To			
CONUS Ship	1.793 (1.780)	1.612 (1.270)	1.460 (1.189)
OCONUS Shore	0.308 (0.209)*	0.440 (0.261)	0.317 (0.188)*
OCONUS Ship	0.043 (0.055)**	0.176 (0.183)*	0.052 (0.062)**
Number of Commands Listed	0.932 (0.074)	1.011 (0.065)	
Mean Rank by Commands	0.942 (0.059)	0.965 (0.053)	
Successful Match (Yes)	3.331 (3.245)	3.692 (3.195)	3.703 (2.418)**
Scramble (Yes)	0.528 (0.475)	0.963 (0.772)	
Scrambled Off List (Yes)	0.459 (0.434)	0.294 (0.250)	0.287 (0.223)
Observations	123	123	123

Notes: 108 observations dropped due to missing values. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 14: Post-Pilot Survey: Logistic Regressions

Regressor	Estimate (std. error)	Estimate (std. error)	Estimate (std. error)	Estimate (std. error)
Male	0.761 (0.418)			
New Graduate (Yes)		2.069 (1.001)		
Officer Rate				
O5 (CDR)			0.214 (0.249)	
O4 (LCDR)			0.580 (0.644)	
O3 (LT)			0.453 (0.514)	
Current Command				
CONUS Ship				0.543 (0.327)
OCONUS Shore				0.453 (0.260)
OCONUS Ship				0.313 (0.301)
Observations	127	127	127	123

Notes: 104 observations dropped due to missing values in the first three model specifications and 108 observations dropped due to missing values in the last model specification. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 15: Post-Pilot Survey: Univariate logistic regressions on demographic regressors

Regressor	Estimate (std. error)	Estimate (std. error)	Estimate (std. error)	Estimate (std. error)
Matched To				
CONUS Ship	0.769 (0.513)			
OCONUS Shore	0.350 (0.183)**			
OCONUS Ship	0.115 (0.098)**			
# of Commands Listed		1.002 (0.052)		
Mean Rank by Commands			0.934 (0.045)	
Rank-Pick-Match				0.838 (0.094)
Observations	127	127	127	114

Notes: 113 observations (Rank-Pick-Match) and 104 observations (all other regressors) dropped due to missing values. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table 16: Post-Pilot Survey: Univariate logistic regressions on regressors derived from pilot project

Glossary

Billet Navy term for an assignment or position. Officers typically rotate to a new billet approximately every three to five years. 126–140, 145, 149–155

CONUS (In the) Continental United States. 126, 127, 131, 133, 135, 137, 138, 140, 145–147, 151

Detailer A Navy officer whose job it is to assign other officers to their next respective billets; also known as an assignments officer. 125, 127–129, 132, 140, 151, 153

New Residency Graduate After graduating from medical school, physicians must complete further training in a residency program before they can practice independently in that specialty. New graduates have just completed residency training and are available for their first assignment as a USN residency-trained physician. 140

OCONUS Outside the Continental United States. 126, 127, 131, 133, 134, 137, 138, 140, 145–147, 149–152, 154

Officer Rate An officer's rank in USN. For physicians, takes values in O3 (lieutenant, LT), O4 (lieutenant commander, LCDR), O5 (commander, CDR), O6 (captain, CAPT). Also known as seniority, officer rank, military rank. 130, 131, 133, 135, 140, 154

PCS Permanent Change of Station; an event in which an officer's assignment to a billet expires, and that officer accepts new orders for a different billet. 152

Retour An event in which an officer's assignment to a billet expires, and that officer accepts new orders for the same billet. 152

Ship billet A billet in direct support of a military unit; examples include positions with ships, infantry, or special forces; also known as an operational billet. 126, 127, 131, 133, 134, 137, 138, 140, 145–147, 150–152

Shore billet A billet in a military clinic or hospital, also known as a military treatment facility (MTF), or in a non-operational, nonclinical role; also known as, and sometimes further categorized as a non-operational or administrative billet. 126, 127, 131, 133, 135, 140, 145–147

References

Davis, J. M., Greenberg, K., & Jones, D. (2023). An experimental evaluation of deferred acceptance: Evidence from over 100 army officer labor markets. *NBER working paper*.

Ferguson, M. D., Hill, R., & Lunday, B. (2020). A scenario-based parametric analysis of the army personnel-to-assignment matching problem. *Journal of Defense Analytics and Logistics*, *4*(1), 89–106.

Fragiadakis, D., Iwasaki, A., Troyan, P., Ueda, S., & Yokoo, M. (2016). Strategyproof matching with minimum quotas. *ACM Transactions on Economics and Computation (TEAC)*, *4*(1), 1–40.

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

Greenberg, K., Pathak, P. A., & Sönmez, T. (2024). Redesigning the u.s. army's branching process: A case study in minimalist market design. *American Economic Review*, *114*(4), 1070-1106.

Klein, T. (2018). Analysis of stable matchings in r: Package matchingmarkets. *The Comprehensive R Archive Network. Retrieved from https://cran. r-project. org*.

Korkmaz, I., Gökçen, H., & Çetinyokuş, T. (2008). An analytic hierarchy process and two-sided matching based decision support system for military personnel assignment. *Information Sciences*, *178*(14), 2915–2927.

Roth, A. E. (1986). On the allocation of residents to rural hospitals: A general property of two-sided matching markets. *Econometrica*, *54*(2), 425–427.

Salyers, M. P., Bonfils, K. A., Luther, L., Firmin, R. L., White, D. A., Adams, E. L., & Rollins, A. L. (2017). The relationship between professional burnout and quality and safety in healthcare: a meta-analysis. *Journal of General Internal Medicine*, *32*(4), 475–482.

Sönmez, T., & Switzer, T. B. (2013). Matching with (branch-of-choice) contracts at the united states military academy. *Econometrica*, *81*(2), 451–488.

USN. (2024). *Navy Military Personnel Manual.* NAVPERS 15560D. Retrieved from https://tinyurl.com/3k5rnnzh (Accessed March 27, 2024)

Yang, W., Giampapa, J., & Sycara, K. (2003). *Two-sided matching for the us navy detailing process with market complication* (Tech. Rep.). Carnegie-Mellon University: Technical Report CMU-RI-TR-03-49, Robotics Institute.

How to prepare a paper for submission

Before submitting a paper to **this Journal**, the authors are advised to follow closely the following instructions to prepare the paper.

1. Papers submitted to this Journal must be unpublished and original work that is neither under review elsewhere nor will be submitted elsewhere for publication without withdrawing from this Journal.
2. This Journal requires that all results (experimental, empirical and computational) be replicable. All underlying data necessary to replicate results must be made available to the Journal.
3. Papers must be submitted in electronic form (preferably as pdf files) and the size of the text font must be at least 12 point. Each figure and table must be included on the relevant page of the paper and should not be collected at the end of the paper. The list of references should appear after any appendices, as the last part of the paper.
4. There is no restriction on the number of pages of any submitted manuscript. We seek to process any first submission within 3 months. However, very long manuscripts may take considerably more time to review.
5. Each submitted paper should have an abstract of no more than 150 words containing no mathematical formulas, complete with no more than 3 suggested keywords and JEL codes. We also encourage authors to make the introduction of their submitted articles understandable to the widest audience possible.
6. The first page of each submitted paper should contain every author's e-mail address, phone number and affiliation.
7. The editors or the publishing Society will not hold any responsibility for views expressed by authors in this Journal.

How to submit a paper

Papers should be submitted electronically in PDF to the Journal of Mechanism and Institution Design through the website `https://www.mechanism-design.org/`.

Aims & scope of the journal

The Journal of Mechanism and Institution Design aims to publish original articles that deal with the issues of designing, improving, analysing and testing economic, financial, political, or social mechanisms and institutions. It welcomes theoretical, empirical, experimental, historical and practical studies. It seeks creative, interesting, rigorous, and logical research and strives for clarity of thought and expression. We particularly encourage less experienced researchers such as recent PhD graduates to submit their work to the Journal and are sympathetic towards those papers that are novel and innovative but which have been unsuccessful elsewhere. We hope that the published articles will be interesting and valuable to a broad audience from the areas of economics, finance, politics, law, computer science, management, history, mathematics, government, and related disciplines. The journal is an open-access, independent, peer-reviewed, non-profit, English-language journal with the purpose of disseminating and sharing the latest knowledge and understanding of the subject widely.

In order for any work that is published by the Journal of Mechanism and Institution Design to be freely accessible to the widest audience possible, when a paper is accepted by this Journal for publication, its author(s) will be asked to release the paper under the Creative Commons Attribution-Non-Commercial license. While the authors retain the copyright of their published work, this license permits anyone to copy and distribute the paper for non-commercial purposes provided that both the author(s) of the article and the Journal are properly acknowledged. For details of the copyrights, please see the “human-readable summary” of the license.