Journal of Mechanism and Institution Design

CONTENTS

Volume 7, Issue 1, December 2022
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
http://www.mechanism-design.org

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

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

Cover & Logo Artwork @ Jasmine Yang
LaTeX Editor & Journal Design @ Paul Schweinzer (using 'confproc')
LaTeX Editorial Assistants @ Theresa Marchetti & Daniel Rehsmann

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

A Letter from the Editor

The year of 2022 has seen more catastrophic events. The global Covid pandemic is not yet over. There immediately came a big war in Europe, drought and famine in Africa, a short-lived UK government, and a political backward movement in China, etc. Energy crises, economic crises, and political crises! One after another. Prices hit record highs. Millions and millions of people have suffered and have been badly hurt! All this cries for better rules, better institutions, better governments, better international organizations, and better design of them.

Despite all the mentioned bad developments on global scales, we still have some good news to share with you about our Society for the Promotion of Mechanism and Institution Design on which we can have some influence. The Society has had a successful Conference on Mechanism and Institution Design at the National University of Singapore, Singapore, July 11–15, 2022. There were 246 scheduled talks and four keynote speeches given by Fuhito Kojima, Dan Kovenock, Alessandro Pavan, and Rakesh Vohra. We wish to express our gratitude to Jingfeng Lu, the organizer, and his colleagues for their time, efforts, and enthusiasm. Our next bi-annual conference will take place in Budapest, Hungary, in the summer of 2024. Corvinus University is the host and Peter Biro is the main organizer. We are looking forward to this event.

Finally, I would like to say a bit more about our Society. It is an independent learned society, a recognized UK charity body, managing the bi-annual Conference on Mechanism and Institution Design and its flagship Journal of Mechanism and Institution Design. Its mission is to advance and promote education and research for the public benefit in the subject of mechanism and institution design. The Journal aims to publish high quality articles in the stated fields and subjects. It is radically different from those journals of commercial publishers in that it is totally free of charge to authors and readers, and provides free open online access to everyone. Our Society fully and independently manages its referring, editing, designing, and production. We rely on voluntary contributions by our Society's members and believe this mode of scientific publishing serves our profession and our society better. We hope more and more people will share our vision and support us.

Zaifu Yang, York, December 3rd, 2022.

Journal of Mechanism and Institution Design
ISSN: 2399-844X (Print), 2399-8458 (Online)
DOI:10.22574/jmid.2022.12.001

OBJECT-BASED UNAWARENESS: THEORY AND APPLICATIONS

Oliver J. Board

Paul | Weiss, USA
ojboard@paulweiss.com

Kim-Sau Chung

Hong Kong Baptist University, China
kschung@hkbu.edu.hk

ABSTRACT

In this paper and its companion paper, Board & Chung (2021), we provide foundations for a model of unawareness that can be used to distinguish between what an agent is unaware of and what she simply does not know. At an informal level, this distinction plays a key role in a number of recent papers such as Tirole (2009) and Chung & Fortnow (2016). Here we provide a set-theoretic (i.e., non-linguistic) version of our framework. We use our *object-based unawareness structures* to investigate two applications. The first application provides a justification for the *contra proferentem doctrine* of contract interpretation, under which ambiguous terms in a contract are construed against the drafter. Our second application examines speculative trade.

Keywords: Unawareness, legal doctrine, no-trade theorem.

JEL Classification Numbers: D83, D86, D91, K12.

This paper was first circulated in 2008. The literature has since grown much bigger than is reflected in our references. We apologize for not being able to do justice to this subsequent literature. We thank Eddie Dekel, Lance Fortnow, Joseph Halpern, Jing Li, Ming Li, and seminar participants at various universities for very helpful comments. We also thank Zaifu Yang for inviting us to submit the paper to this *Journal*. All errors are ours.

1. INTRODUCTION

THERE are two strands of literature on unawareness and it seems that they are unaware of each other.

The first unawareness literature (let's call it the *applied* literature) consists of applied models, such as Tirole (2009) and Chung & Fortnow (2016), where agents are uncertain whether they are aware of everything that their opponents are aware of, and have to strategically interact under these uncertainties. For example, in Tirole (2009), a buyer and a seller negotiate a contract as in the standard hold-up problem. At the time of negotiation, there may or may not exist a better design for the product. Even if a better design exists, however, the contracting parties may not be aware of it. If a party is aware of it, he can choose whether or not to point it out to the other party. But even if he is not aware of it, he *is* aware that a better design may exist and his opponent may be aware of this better design. In Tirole's words, "parties are unaware, but aware that they are unaware"; and they have to negotiate under this uncertainty. Chung & Fortnow (2016) consider the plight of an American founding father drafting a Bill of Rights that will be interpreted by a judge 200 years later. The founding father is aware of some rights, but is uncertain whether or not there are other rights that he is unaware of. Here, as in Tirole (2009), the founding father is unaware, but aware that he may be unaware; and he has to decide how to write the Bill of Rights under this uncertainty.

The second unawareness literature (let's call it the *foundational* literature) attempts to provide a more rigorous account of the properties of unawareness: see, e.g., Fagin & Halpern (1987), Modica & Rustichini (1994), Modica & Rustichini (1999), Dekel et al. (1998), Halpern (2001), Li (2009), Halpern & Rêgo (2006), Sillari (2006), and Heifetz et al. (2006), Heifetz et al. (2013). These authors are motivated by the concern that *ad hoc* applied models, if not set up carefully enough, may go awry in the sense that agents in those models may violate rationality in some way, as captured by various introspection axioms first articulated in Modica & Rustichini (1994) and Dekel et al. (1998) (which we shall refer to as the DLR axioms hereafter).[1] The rest of this literature proposes various models that are set up carefully enough to take these concerns into account.

[1] In particular, two of the key DLR axioms are *KU-introspection* ("the agent cannot know that he is unaware of a specific event") and *AU-introspection* ("if an agent is unaware of an event E, then he must be unaware of being unaware of E").

These two literatures are somewhat disconnected. For example, Tirole makes no reference to any work in the foundational literature, nor does he explain whether or not his agents satisfy the DLR axioms that are the main concerns of that literature. Similarly, none of the studies in the foundational literature explains whether Tirole's model fits in their framework, and if not, whether Tirole's agents violate some or all of the DLR axioms. This paper and Board & Chung (2021) attempt to connect these two literatures.

There is a reason why it is difficult to directly compare Tirole's model with the majority of the models proposed in the foundational literature. To propose a model and to provide foundations for it, an author needs to explain how her model should be interpreted. This is typically done by showing how her model assigns truth conditions to each sentence in a particular formal language; i.e., by the procedure of systematically giving yes/no answers to a laundry list of questions such as: "at state w, does agent i know that it is sunny in New York?"[2] Note, however, the formal language chosen by the author defines the laundry list of questions she is ready to give yes/no answers to. A question not expressible in her chosen formal language is hence not a legitimate question. The answer to it is neither yes nor no—she simply is not ready to say.

Unfortunately, questions such as "at state w, is agent i aware that he is not aware of everything?" are *not* expressible in the formal languages chosen by many authors in the foundational literature (notable exceptions include Halpern & Rêgo (2006) and Sillari (2006), which we shall return to shortly). The formal languages chosen by these authors do not contain quantifiers such as "everything", thus rendering "aware of everything" an inexpressible concept. In other words, while in Tirole's model, "parties are unaware, but aware that they are unaware", it is difficult to tell whether this is also true of the agents in most of the models proposed in the foundational literature. The answer is neither yes nor no—these authors simply are not ready to say.

Several contributions to the foundational literature, mostly coming from logicians and computer scientists, do work with formal languages that contain quantifiers; see, e.g., Halpern & Rêgo (2006) and Sillari (2006). Their proposed models, however, look very different from applied economic models used in, for example, Tirole (2009) and Chung & Fortnow (2016). For

[2] In many of the studies more familiar to economists (see e.g., Li (2009)), although this procedure is not performed explicitly, there is still a clear way to assign truth conditions within an appropriately-specified formal language according to the author's description of her proposed model.

example, in the model proposed by Halpern & Rêgo (2006), there is a *syntactic* awareness function that assigns to every state and every agent a set of sentences in their chosen formal language. The interpretation is that this set is the set of facts that the agent is aware of at that state. This "list of sentences" approach to construct models is very flexible, but may be deemed unhelpful by economists. This may explain why this approach, while not uncommon among logicians, is rarely seen in economics.[3]

In the specific case of Halpern & Rêgo (2006), there is a deeper reason why their proposed model is *not* the same as the models used in the applied literature. Recall that in the latter models, although agents know what they are aware of, they may be uncertain whether or not they are aware of everything. Such uncertainty cannot arise in the model proposed by Halpern & Rêgo (2006), however.[4]

To summarize, while the assumption that "agents are unaware, but are aware that they are unaware" plays a key role in much of the applied literature of unawareness, the foundations of these models remain unclear. We do not know whether agents in these models violate some or all of he DLR axioms that are the main concerns of the foundational literature. This paper and Board & Chung (2021) aim to provide this missing foundation.

In these two papers, we describe a model, or more precisely a class of

[3] To provide an analogy that may help elucidate this comparison, consider the difference between Aumann's information partition model, where a partition of the state space is used to encode an agent's knowledge of events, and a "list of sentences" approach where knowledge is instead modeled by a list of sentences describing exactly what that agent knows.

[4] For readers who are familiar with Halpern & Rêgo (2006), this can be proved formally as follows. Recall the following definition in Halpern & Rêgo (2006): "*Agents know what they are aware of* if, for all agents i and all states s,t such that $(s,t) \in \mathscr{K}_i$ we have that $\mathscr{A}_i(s) = \mathscr{A}_i(t)$." So it suffices to prove that, in any instance of Halpern & Rêgo (2006) structure, if there is a state t such that agent i is uncertain whether or not there is something he is unaware of, then there must be another state s such that $(s,t) \in \mathscr{K}_i$ but $\mathscr{A}_i(s) \neq \mathscr{A}_i(t)$. Let $\alpha = \exists x \neg A_i x$ represent "there is something that agent i is unaware of". Therefore, $\neg\alpha$ means "there is nothing that agent i is unaware of". Let $\beta = A_i\alpha \wedge A_i\neg\alpha \wedge \neg X_i\alpha \wedge \neg X_i\neg\alpha$ represent "agent i is aware of both α and $\neg\alpha$ but he does not know whether α or $\neg\alpha$ is true (recall that X_i is Halpern & Rêgo (2006)'s explicit knowledge operator). In short, β means "agent i is uncertain whether or not there is something he is unaware of". Let M be any instance of Halpern & Rêgo (2006)'s structure, and t is a state such that $(M,t) \models \beta$. Then we have $(M,t) \models \neg K_i\alpha \wedge \neg K_i\neg\alpha$ (recall that K_i is Halpern & Rêgo (2006)'s implicit knowledge operator). Therefore, there exists a state s such that $(t,s) \in \mathscr{K}_i$ and $(M,s) \models \neg\alpha$, and another state s' such that $(t,s') \in \mathscr{K}_i$, and $(M,s') \models \alpha$. Since $\alpha = \exists x \neg A_i x$, there exists ϕ such that $\phi \in \mathscr{A}_i(s)$ and $\phi \notin \mathscr{A}_i(s')$. But that means at least of one $\mathscr{A}_i(s)$ and $\mathscr{A}_i(s')$ is different from $\mathscr{A}_i(t)$.

models, called *object-based unawareness structures* (OBU structures). Readers will find that these structures encompass models used in the applied literature. In comparison with the applied literature, however, we provide complete and rigorous foundations for these structures. The formal language we choose to work with is rich, and in particular contains quantifiers, enabling us to describe explicitly whether or not agents are aware that they are unaware. We provide an axiomatization for these structures and verify that all of the DLR axioms are satisfied. The value of thinking about agents who exhibit this kind of uncertainty has already been demonstrated by the existing applied literature; we demonstrate the tractability of our framework by considering further applications.

A key feature of our structures is that unawareness is object-based: A seller may be unaware of a better design, or a founding father may be unaware of a particular right. In contrast, in models of unforeseen contingencies, agents cannot foresee every contingency, or every state. This raises the question of whether the agents in our structures are aware of every state. We do not have an answer to this question. As we explained above, our understanding of any proposed model is constrained by the formal language we choose to work with. Although we have already chosen to work with a formal language much richer than most in the foundational literature, there are still questions that fall outside of it. We do not have answers to these questions, simply because we do not speak that language.

The division of labor between this paper and Board & Chung (2021) is as follows. In Board & Chung (2021), we give the model-theoretic description of OBU structures by showing how they assign truth conditions to every sentence of a formal language. We then prove a model-theoretic soundness and completeness theorem, which characterizes OBU structures in terms of a system of axioms. We then verify that agents in OBU structures do not violate any of the DLR axioms that are generally considered to be necessary conditions for a plausible notion of unawareness. Board & Chung (2021) also contain a more complete literature review, as well as a discussion of several variants of OBU structures.

In this paper, we give a set-theoretic description of the OBU structures. Although less formal than the model-theoretic treatment, we hope this will be more accessible to the general audience. In parallel to the model-theoretic soundness and completeness theorem in Board & Chung (2021), we prove set-theoretic completeness results in this paper.

The second half of this paper considers two applications. First, we use the model to provide a justification for the *contra proferentem doctrine* of contract interpretation, commonly used to adjudicate ambiguities in insurance contracts. Under *contra proferentem*, ambigous terms in a contract are construed against the drafter. Our main result is that when the drafter (the insurer) has greater awareness than the other party (the insured), *and when the insured is aware of this asymmetry*, *contra proferentem* minimizes the chances that the insured forgoes gain of trade for fear of being exploited. On the other hand, when there is no asymmetric awareness, efficiency considerations suggest no reason to prefer *contra proferentem* over an alternative interpretive doctrine that resolves ambiguity in favor of the drafter.

From the perspective of our theory, an argument common among legal scholars as far back as Francis Bacon, that *contra proferentem* encourages the insurer to write clearer contracts, misses the point. If a more precise contract increases the surplus to be shared between the insurer and the insured, market forces provide incentives to draft such a contract regardless of the interpretive doctrine employed by the court. The advantage of *contra proferentem* is rather that it enables the insurer to draft more acceptable contracts, by expanding the set of events that he can credibly insure.

Our second application examines speculative trade. We first generalize the classical No Trade Theorem to situations where agents are delusional but nevertheless act so as to satisfy a weaker condition called terminal partitionality. We then introduce the concepts of *living in denial* (i.e., agents believe, perhaps incorrectly, that there is nothing that they are unaware of) and *living in paranoia* (i.e., agents believe, perhaps incorrectly, that there is something that they are unaware of). We show that both living in denial and living in paranoid, in the absence of other forms of delusion, imply terminal partitionality, and hence the no trade theorem result obtains.

The structure of this paper is as follows. Section 2 describes our OBU structures, and Section 3 shows how to incorporate probabilities. Section 4 presents the first application, and Section 5 the second. Section 6 concludes.

2. OBU STRUCTURES

In thi section we introduce OBU structures and present set-theoretic completeness results[5] that provide a precise characterization of the properties of knowledge, unawareness etc. For the sake of transparency, and to aid interpretation, we also include in Appendix A the model-theoretic description of these structures; i.e., we show how OBU structures assign truth conditions for a formal language (a version of first-order modal logic).

2.1. Modeling knowledge and unawareness

An *OBU structure* for n agents is a tuple $\langle W, O, \{O_w\}, \{\mathscr{I}_i\}, \{\mathscr{A}_i\}\rangle$, where:

- W is a set of states;
- O is a set of objects;
- $O_w \subseteq O$ is the set of objects that really exist at state w;
- $\mathscr{I}_i : W \to 2^W$ is an *information function* for agent i; and
- $\mathscr{A}_i : W \to 2^O$ is an *awareness function* for agent i.

Intuitively, $\mathscr{I}_i(w)$ indicates the states that agent i considers possible when the true state is w, while $\mathscr{A}_i(w)$ indicates the objects she is aware of. The sets O_w will not be used until we describe quantified events in section 2.3 below.

In the standard information partition model familiar to economists, events are represented as subsets of the state space, corresponding to the set of states in which some given proposition is true. In OBU structures, we try to carry around one more piece of information when we represent an event, namely the set of objects referred to in the verbal description of that event. Formally, an event is an ordered pair (R,S), where $R \subseteq 2^W$ is a set of states and $S \subseteq 2^O$ is a set of objects; we call R the *reference* of the event (denoted by $ref(R,S)$), corresponding (as before) to the set of states in which the proposition is true; and S is the *sense* of the event (denoted by $sen(R,S)$), listing the set of objects referred to in the proposition. (To give an example, the events representing the propositions "the dog barked" and "the dog barked and the cat either did

[5] This purely semantic approach to epistemic logic was pioneered by Halpern (1999).

or did not meow" have the same reference but difference senses.) We sometimes abuse notation and write (R,a) instead of $(R,\{a\})$, and (w,S) instead of $(\{w\},S)$. We use $\mathscr{E}$ to denote the set of all events, with generic element E.

We now define two operators on events, corresponding to "not" and "and":

$$\begin{aligned} \neg(R,S) &= (W \setminus R, S), \\ \wedge_j (R_j,S_j) &= \left(\cap_j R_j, \cup_j S_j\right). \end{aligned}$$

The negation of an event holds at precisely those states at which the event does not hold, but it refers to the same set of objects. The conjunction of several events holds only at those states at which *all* of those events hold, and it refers to each set of objects. It will often be convenient to use disjunction ("or") as well, defined in terms of negation and conjunction as follows:

$$\begin{aligned} \vee_j (R_j,S_j) &= \neg\left(\wedge_j \neg (R_j,S_j)\right) \\ &= \left(\cup_j R_j, \cup_j S_j\right). \end{aligned}$$

In OBU structures, there are three modal operators for each agent, representing awareness, implicit knowledge, and explicit knowledge:

$$\mathsf{A}_i(R,S) = (\{w \mid S \subseteq \mathscr{A}_i(w)\},S) \text{ (awareness)} \quad (1)$$

$$\mathsf{L}_i(R,S) = (\{w \mid \mathscr{I}_i(w) \subseteq R\},S) \text{ (implicit knowledge)} \quad (2)$$

$$\mathsf{K}_i(R,S) = \mathsf{A}_i(R,S) \wedge \mathsf{L}_i(R,S) \text{ (explicit knowlege)} \quad (3)$$

Intuitively, an agent is aware of an event at w if she is aware of every object in the *sense* of the event; and the agent implicitly knows an event at state w if the *reference* of the event includes every state she considers possible. However, implicit knowledge is not the same as explicit knowledge, and the latter is our ultimate concern. Implicit knowledge is merely a benchmark that serves as an intermediate step to modeling what an agent actually knows. Intuitively, an agent does not actually (i.e., explicitly) know an event unless he is aware of the event *and* he implicitly knows the event. Notice that A_i, L_i, and K_i do not change the set of objects being referred to.

It is easy to verify that awareness and implicit knowledge satisfy the following properties (where we suppress the agent-subscripts):

A1 $\wedge_j \mathsf{A}\left(R,S_j\right) = \mathsf{A}\left(R,\cup_j S_j\right)$

A2 $\mathsf{A}(R,S) = \mathsf{A}(R',S)$ for all R,R'

A3 $\mathsf{A}(R,\varnothing) = (W,\varnothing)$

A4 $\mathsf{A}(R,X) = (R',X)$ for some R'

L1 $\mathsf{L}(W,O) = (W,O)$

L2 $\wedge_j \mathsf{L}\left(R_j,S\right) = \mathsf{L}\left(\cap_j R_j,S\right)$

L3 $\mathsf{L}(R,S) = (R',S)$ for some R'

L4 if $\mathsf{L}(R,S) = (R',S)$ then $\mathsf{L}(R,S') = (R',S')$

The following results show that L1–L4 and A1–A4 also provide a precise characterization of awareness and implicit knowledge, respectively.

Proposition 1. *Suppose that* A_i *is defined as in (1). Then:*

1. A_i *satisfies A1–A4; and*
2. *if* A'_i *is an operator on events which satisfies A1–A4, we can find an awareness function* $\mathscr{A}_i$ *such that* A'_i *and* A_i *coincide.*

Proposition 2. *Suppose that* L_i *is defined as in (2). Then:*

1. L_i *satisfies L1–L4; and*
2. *if* L'_i *is an operator on events which satisfies L1–L4, we can find an information function* $\mathscr{I}_i$ *such that* L'_i *and* L_i *coincide.*

The proofs of these and all other results can be found in the appendix.

2.2. Introducing Properties

In an OBU structure, we take as primitives not individual events such as "John is tall", but rather individual *properties* such as "... is tall". Intuitively, the property "... is tall" can be thought of as a correspondence from objects to states, telling us for each object at which states it possesses this property. More generally, properties can be represented as functions from objects to events: $p : O \to \mathscr{E}$ such that

$$p(a) = (R^p_a, S^p \cup \{a\}) \text{ for some } R^p_a \subseteq W \text{ and some } S^p \subseteq O.$$

Intuitively, R_a^p is the set of states where object a possesses property p, and S^p is the set of objects referred to in the description of the property; for example, if p is the property "... is taller than Jim", then $S^p = \{Jim\}$. Note that S^p could be the empty set, for example if p is the property "... is tall". Let $\mathscr{P}$ denote the class of all these functions.

REMARK: In many applications, such as the one we will study in Section 4, the set of properties that are relevant to the problem at hand is a much smaller set than $\mathscr{P}$, and hence not every (R,S) pair is a representation of a proposition like "John is tall".

REMARK: Although we have only described 1-place properties, this is without loss of generality, because we can build up n-place properties from n 1-place properties. Suppose we want to construct the 2-place property $taller(a,b)$, to be interpreted as "a is taller than b". We start with a family of 1-place properties $\{p_a : O \to \mathscr{E}\}_{a \in O}$, to be interpreted "$a$ is taller than ...". Define $f : O \to \mathscr{P}$ as $f(a) = p_a$. Then the two-place property $taller : O^2 \to \mathscr{E}$ is defined by taller $(a,b) = f(a)(b)$. Notice that, in particular, the sense of the event $taller(a,b)$ is $\{a,b\}$, because

$$sen(f(a)(b)) = S^{f(a)} \cup \{b\} = \{a\} \cup \{b\}.$$

We can also take negations, conjunctions, and disjunctions of properties:

$$\begin{aligned} \neg p &: O \to \mathscr{E} \text{ such that } (\neg p)(a) = \neg(p(a)) \\ p \wedge q &: O \to \mathscr{E} \text{ such that } (p \wedge q)(a) = p(a) \wedge q(a) \\ p \vee q &: O \to \mathscr{E} \text{ such that } (p \vee q)(a) = p(a) \vee q(a) \end{aligned}$$

We also use $p \to q$ as shorthand for $\neg p \vee q$.

REMARK: It is worth noting that the concept of negation defined above does not coincide with the everyday English notion of "opposites" (as in "short is the opposite of tall"). There are two reasons for this: first, even if we restrict attention to people (humans), we might argue some people are neither tall nor short (for instance, an white male who is 5 foot tall); second, there are objects which are neither tall nor short simple because they don't have a height at all (for instance, an abstract object such as "a thought". Therefore we prefer to think of tall and short as two separate properties, allowing for the possibility that short is not the same as not tall.

2.3. Quantified Events

In many applications, we want to deal not only with events such as "a is a better design" and "agent i knows that a is a better design", but also events such as "agent i is not aware of any better design" and "agent i does not know whether there is a better design that he is unaware of". These events involve *quantification*. In this section, we show how they are handled in OBU structures.

To begin with, we should note that everyday English admits multiple interpretations of quantifiers (such as the word "all"), corresponding to different scopes implicit in the conversation: the "universe of objects" referred to by the word "all" can vary. We often freely switch back and forth among different interpretations, without making the scope explicit, and leaving it for the context to resolve the ambiguity. In a formal model, however, these different interpretations must be explicitly distinguished by different quantifiers. Two particular quantifiers that may get confused are the *possibilitist quantifier* and the *actualist quantifier*; the former has a scope that spans all possible objects, while the latter has a scope that spans only those objects that really exist at a given state. The quantifier that is used in OBU structures is the actualist one.

To illustrate the difference between these two quantifiers, consider the following application. Suppose we want to model Hillary's uncertainty regarding whether or not Bill has an illegitimate child. The simplest way to do it is to have Hillary consider as possible two different states, w_1 and w_2, but Bill's illegitimate child really exists at only one of these states. Using a to denote "Bill's illegitimate child", it means $a \in O_{w_1} \subset O$ but $a \notin O_{w_2}$. Since Hillary cannot tell apart these two states, she does not know for sure whether Bill has an illegitimate child or not. However, such a simple model of Hillary's uncertainty "works" only because the existential quantifier used by this simple model is the actualist one. If a reader misinterprets the model as using the possibilitist quantifier, he would have regarded it as a poor model of Hillary's uncertainty: "Since Bill's illegitimate child 'exists' at every state that Hillary considers possible, Hillary knows for sure that Bill has an illegitimate child, and hence there is no uncertainty at all!"

We define possibilitist-quantified events first, because they are simpler, and can be used as an intermediate step to define actualist-quantified events. For any property $p \in \mathscr{P}$, let $\overline{\mathsf{All}}\, p$ denote the event that "all objects satisfy property p", where "all" is interpreted in the possibilitist sense. Formally, $\overline{\mathsf{All}}$

is a mapping from properties to events, such that

$$\overline{\mathsf{All}}\, p = \left(\cap_{a\in O} R_a^p, S^p\right).$$

So $\overline{\mathsf{All}}\, p$ holds at precisely those worlds where $p(a)$ is true for each objects a in the universal set O, and it refers only to those objects referred to by property p.

We defined actualist-quantified events, or simply *quantified events*. First recall that an OBU structure specifies, for each state w, the set $O_w \subseteq O$ of objects that really exist at that state. We define a special property *re* ("... is real") in terms of these sets:

$$re(a) = \left(\{w \mid a \in O_w\}, a\right). \tag{4}$$

Let $\mathsf{All}\, p$ denote the event that "all objects satisfy property p", where "all" is interpreted in the actualist sense. Formally, All is a mapping from properties to events, such that

$$\mathsf{All}\, p = \left(\cap_{a\in O} R_a^{re\to p}, S^p\right). \tag{5}$$

Intuitively, $\mathsf{All}\, p$ holds at every state where all real objects possess property p; and the sense of $\mathsf{All}\, p$ is precisely the objects used to describe property p. It is easy to verify that the actualist quantifier satisfies the following properties:

All**1** $\mathsf{All}\left(\wedge_j p_j\right) = \wedge_j \left(\mathsf{All}\, p_j\right)$

All**2** if $w \in R_a^p$ for every $a \in O$, then $w \in ref\left(\mathsf{All}\, p\right)$

All**3** if $R_a^p = R_a^q$ for every $a \in O$, then $ref\left(\mathsf{All}\, p\right) = ref\left(\mathsf{All}\, p\right)$

All**4** $sen\left(\mathsf{All}\, p\right) = S^p$

The following result shows that All1 – All4 also provide a precise charactization of the actualist quantifier.

Proposition 3. *Suppose that* All *is defined as in (4) and (5). Then:*

1. All *satisfies* All1 – All4*; and*
2. *if* All' *is a mapping from properties to events which satisfies* All1 – All4*, we can find a collection of real objects* $\{O_w\}$ *such that* All' *and* All *coincide.*

3. OBU STRUCTURES WITH PROBABILITIES

It is easy to introduce probabilistic beliefs into the OBU structures, although Board & Chung (2021)'s axiomatization does not include this part. We first introduce implicit beliefs, once again as a benchmark case that serves as an intermediate tool to modeling what the agent actually believes. The relation between explicit beliefs (i.e., an agent's actual beliefs) and implicit beliefs is then analogous to the relation between explicit knowledge and implicit knowledge.

Let us begin with an OBU structure $\langle W, O, \{O_w\}, \{\mathscr{I}_i\}, \{\mathscr{A}_i\}\rangle$. To avoid unnecessary complications, let's assume that W is finite. Augment the OBU structure with $\{q_i\}_{i\in N}$, where each q_i is a probability assignment that associates with each state w a probability distribution on W satisfying $q_i(w)(\mathscr{I}_i(w))) = 1$ (i.e., an agent (implicitly) assigns probability 1 to those states that he considers possible when the true state is w). For any real number r, we introduce two belief operators for each agent, mapping any given event $E = (R,S) \in \mathscr{E}$ to the events that an agent implicitly and explicitly, respectively, believes that E holds with probability at least r:

$$\overline{\mathsf{B}}_i^r(R,S) = (\{w \mid q_i(w)(R) \geq r\}, S) \text{ (implicit belief)} \quad (6)$$

$$\mathsf{B}_i^r(R,S) = \mathsf{A}_i(R,S) \wedge \overline{\mathsf{B}}_i^r(R,S) \text{ (explicit belief).} \quad (7)$$

An *augmented OBU structure* is a tuple $\langle W, O, \{O_w\}, \{\mathscr{I}_i\}, \{\mathscr{A}_i\}, \{q_i\}\rangle$.

The common prior assumption is considered controversial, even in the absence of unawareness (Morris, 1995; Gul, 1998). Nevertheless, to facilitate comparison with the existing literature in Section 5, we introduce it here. We say that an augmented OBU structure satisfies the *common prior assumption* if there exists a probability distribution q on W such that, whenever $q(\mathscr{I}_i(w)) > 0$, we have

$$q_i(w)(\cdot) = q(\cdot \mid \mathscr{I}_i(w)),$$

where $q(\cdot \mid \mathscr{I}_i(w))$ is the conditional probability distribution on W given $\mathscr{I}_i(w)$. When an augmented OBU structure satisfies the common prior assumption, we can represent it as the tuple $\langle W, O, \{O_w\}, \{\mathscr{I}_i\}, \{\mathscr{A}_i\}, q\rangle$, and simply call it an OBU structure with common prior.

4. THE CONTRA PROFERENTEM DOCTRINE

Verba fortius accipiuntur contra proferentem (literally, "words are to be taken most strongly against him who uses them") is a rule of contractual interpretation which states that ambiguities[6] in a contract should be construed against the party who drafted the contract. This rule (henceforth *cp doctrine*) finds clear expression in the *First Restatement of Contracts*[7] (1932) as follows:

> Where words or other manifestations of intention bear more than one reasonable meaning an interpretation is preferred which operates more strongly against the party from whom they proceed, unless their use by him is prescribed by law.

Although the principles for resolving ambiguity are more nuanced in the *Second Restatement* (1979), the cp doctrine is widely applied in the context of insurance contracts; indeed, Abraham (1996) describes it as "the first principle of insurance law".

In this section, we use OBU structures to formalize the rationale behind this rule. In particular, we compare it with the opposite doctrine that resolves ambiguity *in favor of* the drafter. We first show that there is a form of symmetry between these two doctrines, and neither systematically outperforms the other if there is no asymmetric unawareness. We then introduce asymmetric unawareness and explain in what sense the cp doctrine is a superior interpretive doctrine.

Let an OBU structure with common prior $\langle W, O, \{O_w\}, \{\mathscr{I}_i\}, \{\mathscr{A}_i\}, q\rangle$ be given.[8] Assume that there are two agents. Agent 1 is a (female) risk-neutral

[6] "Ambiguity" is an ambiguous term in economics, and often refers to situations where decision makers entertain multiple prior probability distributions. Here, we are referring to the layman's use of the word, that is to a situation where language is susceptible to multiple interpretations.

[7] The *Restatements of the Law* are treatises published by the American Law Institute as scholarly refinements of black-letter law, to "address uncertainty in the law through a restatement of basic legal subjects that would tell judges and lawyers what the law was." Although nonbinding, the authoritativeness of the Restatements is evidenced by their near-universal acceptance by courts throughout the United States.

[8] Given our earlier comments about the common prior assumption, the reader may wonder why we impose this assumption here. The common prior assumption allows us to state our results neatly. But we otherwise do not believe that the comparison between different doctrines depends on this assumption.

insurer and agent 2 is a (male) risk-averse insured. In the absence of any insurance contract between the agents, agent 1's income is \$0 in every world, while agent 2's income is \$0 in some worlds and \$1 in other worlds. We can think of 0 income as the result of some negative income shock, which the risk-averse agent 2 would like to insure against. Agent 1's utility is equal to her income, and agent 2's utility is $U(\cdot)$, which is strictly increasing and strictly concave in his income.

One of the elements in O, denoted by ι, is agent 2's income. (We will explain what else is contained in O later.) Let $Z \subsetneq W$ be the (nonempty) set of states in which agent 2 suffers an income shock. The event "agent 2 suffers an income shock" is hence $E = (Z, \iota)$. It is natural to assume that agent 2 is always aware of his own income (i.e., $\iota \in \mathscr{A}_2(w)$ for every w), and so agent 2 can always form an explicit probabilistic belief about event E (given by $q(ref(E))$=$q(Z)$).

To make the setup as noncontroversial as possible, we make a couple of standard assumptions:

1. Each agent i's $\mathscr{I}_i$ forms a partition of the state space W; i.e., $w \in \mathscr{I}_i(w)$ for every $w \in W$, and $w' \in \mathscr{I}_i(w)$ implies $\mathscr{I}_i(w') = \mathscr{I}_i(w)$.

2. Each agent i (implicitly) knows what he is aware of; i.e., $w' \in \mathscr{I}_i(w)$ implies $\mathscr{A}_i(w') = \mathscr{A}_i(w)$.

We also make an additional assumption motivated by the current application:

3. Agent 1 is aware of more objects than agent 2 is: $\mathscr{A}_2(w) \subseteq \mathscr{A}_1(w)$ for every $w \in W$.

The third assumption captures the idea that agent 1 (the insurer) is the more sophisticated party in this transaction. In what follows we analyze a special case that satisfies these assumptions: agent 1 is aware of everything while agent 2 is aware of nothing except his own income: $\mathscr{A}_1(w) = O$ and $\mathscr{A}_2(w) = \{\iota\}$ for all w; and both agents are completely uninformed: $\mathscr{I}_i(w) = W$ for all w and $i = 1, 2$. This allows us to abstract away from the classical adverse selection problem, which is already well understood, and focus instead on the interaction between contractual ambiguity and asymmetric awareness.

Note that, although we make the extreme assumption that agent 2 is aware of nothing (except his own income), we do not preclude that he is aware of his

own unawareness. For example, as long as $O_w \setminus \{\iota\} \neq \varnothing$ for all w, the event "agent 1 is aware of something that agent 2 is unaware of" (where "some" is interpreted in the actualist sense) is the event $(W, \varnothing)$. Since

$$\mathsf{K}_2(W,\varnothing) = \mathsf{A}_2(W,\varnothing) \wedge \mathsf{L}_2(W,\varnothing) = (W,\varnothing) \wedge (W,\varnothing) = (W,\varnothing), \qquad (8)$$

agent 2 *explicitly* knows that "agent 1 is aware of something that agent 2 is unaware of" in every state w.

If we further assume that $O_w = \hat{O} \subset O$ for all w, then agent 2 knows how many objects there are that agent 1 is aware of but agent 2 is not. Although this assumption is not realistic (even if the insured is certain that there are *some* objects that he is unaware of, he will typically be uncertain about the exact number of such objects), it simplifies the analysis considerably. In this preliminary investigation of the cp doctrine, therefore, we add this assumption. To further simplify, we assume that $\hat{O} = O$ until section 4.3.2 where it becomes important to distinguish the two sets.

The timing of the contracting game is as follows. In stage one, agent 1 proposes an insurance contract. The contract specifies a premium, a payment, and the circumstances under which agent 1 (the insurer) has to pay the insurance payment to agent 2 (the insured). A critical assumption is that the payout circumstances have to be described in an exogenously given language, to be defined shortly, and cannot make reference to agent 2's income. Without this assumption, the insurance problem would be trivial. This assumption makes sense when, for example, agent 2's income is not verifiable and hence not contractible, or if contracting on income would create a serious moral hazard problem. In stage two, agent 2 either accepts the contract and pays the premium, or rejects it. If he accepts, we move to stage three, the contract enforcement stage, where nature randomly picks a state according to the probability distribution q, and agent 1 has to pay agent 2 the insurance payment unless she can prove to a court that the payout circumstances do not obtain.

4.1. Contracts and Interpretations

We now define the *contractual language*, which is built up from the following elements (the *vocabulary*):

- $a, b, c \ldots$ — an exogenously given, nonempty list of (names of) objects, which together with agent 2's income ι form the the set O in our OBU structure (i.e., $O \setminus \{\iota\} = \{a, b, c, \ldots\}$).

- $P_1, P_2, \ldots$ — an exogenously given, nonempty list of *predicates*, each of which will later on be construed (by the court) as corresponding to a specific property.[9]

- $\neg$ (not), $\wedge$ (and), $\vee$ (or) — Boolean operators.

Note that by identifying the set of objects' names with the objects themselves, we are assuming that there is no ambiguity in the interpretation of these names; we make the simplifying assumption that all contractual ambiguity relates to which properties the various predicates stand for.

Formally, the contractual language is a collection of sentences, each of which is a finite string of *letters* (i.e., elements of the vocabulary) satisfying a certain grammatical structure. We define this collection recursively as follows:

(i) for each object a and predicate P, $P(a)$ (to be interpreted as "object a is P") is a sentence;

(ii) if ϕ and ψ are sentences, then $\neg\phi$, $\phi \wedge \psi$, and $\phi \vee \psi$ are sentences.

The contractual language, denoted by $\mathscr{L}$, is the smallest set satisfying (i) and (ii).[10] If b and r are objects and F and L are predicates, an example of a sentence in $\mathscr{L}$ is $F(b) \wedge L(r)$, with a possible interpretation of "the basement is flooded and the roof is leaking".

An *insurance contract* is a triple (g, h, ϕ), where $g \in \mathbb{R}_+$ is the insurance premium that agent 2 pays agent 1 *ex ante*, and $\phi \in \mathscr{L}$ is a sentence that describes the circumstances under which agent 1 pays $h \in \mathbb{R}_+$ to agent 2 *ex post*.

Although a predicate P (in the vocabulary of the contractual language) is supposed to correspond to a specific property, whether an object satisfies that property or not is often ambiguous *ex post*. For example, consider a health insurance contract that covers the cost of a hip replacement just when it is medically necessary. Is a patient who is able to walk, but only with a great deal of pain, covered? Some people might say yes, while others would say no.

[9] Without loss of generality, we assume that all these predicates are 1-place. See Section 2 for discussion.

[10] We could further expand our contractual language to include quantifiers. We conjecture that this would not affect our main results.

Without this kind of ambiguity, the cp doctrine would be moot. So we now introduce this kind of ambiguity into our model.

We capture this kind of ambiguity by supposing that there may be disagreement about which property (in an OBU structure) a given predicate corresponds to. Formally, an *interpretation* is a mapping l from predicates to properties. To keep things simple, imagine that there are two sub-populations of society, and each has its own interpretation of every predicate P. Let l_1 and l_2 denote these two interpretations. It is natural to assume that $S^{l_1(P)} = S^{l_2(P)}$.

An interpretation l that maps predicates to properties can be extended to a mapping from the contractual language $\mathscr{L}$ to events in the obvious way:

I1 $l(P(a)) = \left(R_a^{l(P)}, S^{l(P)} \cup \{a\}\right)$;

I2 $l(\neg\phi) = \neg l(\phi)$;

I3 $l(\phi \wedge \psi) = l(\phi) \wedge l(\psi)$;

I4 $l(\phi \vee \psi) = l(\phi) \vee l(\psi)$.

We can now formalize the cp doctrine. The cp doctrine instructs the court to resolve any ambiguity against the party who drafted the contract (i.e., agent 1 in this model). In the example above, if the hip replacement is medically necessary given one interpretation but not the other, then under cp doctrine the court should rule in favor of agent 2 and require agent 1 to payout. Formally, the cp doctrine is a mapping from $\mathscr{L}$ to events given by

$$d_{cp}(\phi) = l_1(\phi) \vee l_2(\phi) \quad \text{for all } \phi \in \mathscr{L}.$$

Note that d_{cp} is *not* an interpretation, since it may not satisfy I2 or I3.

For sake of comparison, we set up a strawman and define the mirror image of the cp doctrine, the *anti-cp doctrine*, which instructs the court to resolve any ambiguity in favor of agent 1. Formally, $d_{anti-cp}$ is given by

$$d_{anti-cp}(\phi) = l_1(\phi) \wedge l_2(\phi) \quad \text{for all } \phi \in \mathscr{L}.$$

The interpretive doctrine of the court is commonly known. Given this interpretive doctrine d, agent 1's problem in stage three (the contract enforcement stage) is to prove to the court that the payout circumstances do not obtain, or equivalently that event $d(\phi)$ has not happened.

We assume that, once the true state w is realized, agent 1 has sufficient evidence to prove that object a satisfies property p *if and only if* (1) a is real ($a \in O_w$), and (2) a does in fact satisfy property p ($w \in R_a^P$). Under our earlier simplifying assumption that $O_w = \hat{O} = O$ for every w, condition (1) is always satisfied.

Finally, we need to explain how agent 2 evaluates a given contract and makes his accept/reject decision accordingly in stage two. This can be tricky, as it depends on how agent 2's awareness changes after he reads the contract (which may mention objects that agent 2 was unaware of before he read it). We postpone this discussion to section 4.3 below, and first consider a benchmark case where there is symmetric awareness between the two agents. The central message from the benchmark case is this: linguistic ambiguity alone (without asymmetric unawareness) is not sufficient to justify the cp doctrine.

EXAMPLE: Let's use an example to illustrate our setup. Consider the simplest case where there is only one object name, a, and one predicate, P, in the contractual language. One can think of a as "the basement", and P as "... is flooded". Suppose there are only two states: w_1 and w_2. At w_1, there is a lot of water in the basement, and everyone in the society would agree that the basement is flooded. But at w_2, the basement is merely wet, and not everyone in the society would think that it is flooded. Therefore we have $l_1(P(a)) = (\{w_1, w_2\}, a)$ and $l_2(P(a)) = (\{w_1\}, a)$. Suppose the contract says that the insured will be compensated when the basement is flooded; i.e., the contract takes the form of $(g, h, P(a))$. Under the cp-doctrine, the insured will be compensated at both states; whereas under the anti-cp doctrine, he will be compensated only at state w_1. As another example, suppose the contract says that the insured will be compensated when the basement is *not* flooded; i.e., the contract takes the form of $(g, h, \neg P(a))$. Under the cp-doctrine, the insured will be compensated at state w_2; whereas under the anti-cp doctrine, he will never be compensated.

4.2. Benchmark: Symmetric Awareness

Before we continue the description of our model, let's first consider the benchmark case of *symmetric awareness*, where $O_1(w) = O_2(w) = O$ for every $w \in W$. In this case, agent 2 is aware of every object that agent 1 is aware of. Since both agents are aware of every object, implicit knowledge/beliefs and explicit knowledge/beliefs coincide. This reduces our model back to a

standard exercise in contract theory. The introduction of an exogenous contractual language does not pose a new methodological challenge, because its only effect is to restrict the contracting parties' ability to approximate a first-best contract. Different interpretive doctrines imply different restrictions on the contracting parties. However, as we shall see shortly, there is a strong symmetry between the restrictions implied by the cp doctrine and those implied by the anti-cp doctrine, and hence no systematic advantage for the former over the latter.

A first best contract is any contract that requires the insurer to pay \$1 to the insured exactly in those states where he suffers an income shock.[11] Recall that Z denotes the set of states where the insured suffers an income shock. Since the contracting parties cannot write contracts that directly refer to agent 2's income, they have to look for (contractible) events that correlate with agent 2's income shock. In other words, they have to look for a $\phi \in \mathscr{L}$ such that, under a given interpretive doctrine d, the set $ref(d(\phi))$ approximates Z. How well $ref(d(\phi))$ approximates Z depends on the prior probability q; or, more precisely, on $q(ref(d(\phi)) \setminus Z)$ and $q(Z \setminus ref(d(\phi)))$.

To make this more precise, let $\mathscr{R}_{cp} = \{ref(d_{cp}(\phi)) \mid \phi \in \mathscr{L}\}$ denote the set of references that can be described under the cp doctrine; similarly, let $\mathscr{R}_{anti-cp} = \{ref(d_{anti-cp}(\phi)) \mid \phi \in \mathscr{L}\}$. Then say that the cp doctrine *systematically out-performs* the anti-cp doctrine if and only if $\mathscr{R}_{anti-cp} \subsetneq \mathscr{R}_{cp}$.

To see that this definition captures the correct intuition, suppose first that $\mathscr{R}_{anti-cp} \not\subseteq \mathscr{R}_{cp}$. Then there is some (non-empty)[12] $R \in \mathscr{R}_{anti-cp} \setminus \mathscr{R}_{cp}$. If $Z = R$ and q is the uniform prior, then full insurance is possible only under the anti-cp doctrine. On the other hand, if $\mathscr{R}_{anti-cp} \subsetneq \mathscr{R}_{cp}$, any insurance outcome achievable under the anti-cp doctrine can be replicated under the cp doctrine, while we can find a case where full insurance is possible only under the cp doctrine.

EXAMPLE CONTINUED: Let's use our earlier example to illustrate what is at stake when the society chooses between the two doctrines. In that example,

$$\mathscr{R}_{cp} = \{\varnothing, \{w_2\}, \{w_1, w_2\}\}.$$

Note that the singleton set $\{w_1\}$ is not in $\mathscr{R}_{cp}$. Therefore, full insurance is not always possible under the cp doctrine. In particular, if $Z = \{w_1\}$ (i.e., the in-

[11] The insurance premium is a pure transfer and hence has no efficiency implications.

[12] It is easy to see that $\varnothing \in \mathscr{R}_{anti-cp} \cap \mathscr{R}_{cp}$.

sured's wealth drop is correlated with how severely his basement is flooded), the contractual language would be found inadequate for the purpose of providing insurance—in fact, the optimal insurance contract will be no insurance in such an unfortunate case. Now, consider the counterfactual case where the parties anticipate that the court would interpret their contract using the anti-cp doctrine. Under such anticipation, they can sign a contract of the form $(g,h,P(a))$; and with $d_{anti-cp}(P(a)) = (\{w_1\},a) = (Z,a)$, perfect insurance can be achieved. But does it mean that the anti-cp doctrine is better than the cp doctrine? The answer is no, because by a symmetric argument we can see that, in case $Z = \{w_2\}$, perfect insurance can be achieved under the cp doctrine but not under the anti-cp doctrine. Without further information regarding which case is more likely, it is impossible to rank the two doctrines.

The following proposition says that $|\mathscr{R}_{anti-cp}| = |\mathscr{R}_{cp}|$, and so it cannot be the case that the cp doctrine systematically outperforms the anti-cp doctrine.

Proposition 4. $|\mathscr{R}_{anti-cp}| = |\mathscr{R}_{cp}|$.

Proof. It suffices to show that $R \in \mathscr{R}_{anti-cp}$ if and only if $W \setminus R \in \mathscr{R}_{cp}$. Suppose $R \in \mathscr{R}_{anti-cp}$. Then there exists $\phi \in \mathscr{L}$ such that $ref(d_{anti-cp}(\phi)) = R$. But $\phi \in \mathscr{L}$ implies $\neg\phi \in \mathscr{L}$. Since $ref(d_{cp}(\neg\phi)) = ref(l_1(\neg\phi) \vee l_2(\neg\phi)) = ref(\neg l_1(\phi) \vee \neg l_2(\phi)) = ref(\neg l_1(\phi)) \cup ref(\neg l_2(\phi)) = (W \setminus ref(l_1(\phi))) \cup (W \setminus ref(l_2(\phi))) = W \setminus (ref(l_1(\phi)) \cap ref(l_2(\phi))) = W \setminus ref(l_1(\phi) \wedge l_2(\phi)) = W \setminus ref(d_{anti-cp}(\phi)) = W \setminus R$, we have $W \setminus R \in \mathscr{R}_{cp}$. The other direction is similar. □

We find it illuminating to contrast Proposition 4 with an argument common among legal scholars as far back as Francis Bacon, that the advantage of *contra proferentem* is to provide incentives for the insurer to write precise contracts. Carolina Care Plan Incorporated v. McKenzie (2007) provides a succinct statement of this argument in a recent ruling: "Construing ambiguity against the drafter encourages administrator-insurers to write clear plans that can be predictably applied to individual claims, countering the temptation to boost profits by drafting ambiguous policies and construing them against claimants." However, in light of Propositon 4, this argument misses the point. According to our theory, more precise contracts will be rewarded by higher premiums regardless of the interpretative doctrine employed by the court.

4.3. Asymmetric Awareness

We now return to the case of asymmetric awareness: $\mathscr{A}_1(w) = O$ and $\mathscr{A}_2(w) = \{\iota\}$ for all $w \in W$. Here, an important modelling question to address is: how would agent 2's awareness changes after he reads a contract which mentions objects that he was previously unaware of?

If agent 2 was unaware of those objects because they slipped his mind, then it would be natural to assume that he becomes aware of them once he reads about them in the contract. If, instead, he was unaware of them because he genuinely had no idea what they were, then it would be more natural to assume that his awareness would not change even after reading the contract. In reality there would likely be some objects in each category, which begs a richer model that distinguishes a slip-the-mind object from a genuinely-clueless object. For the sake of simplicity, we keep the two cases distinct and analyze each in turn.

Although the the slip-the-mind case is not the only case where unawareness can arise, it is the only case that has been considered by other authors so far.[13] However, in the current setup, it turns out that the slip-the-mind case and the benchmark case (with symmetric awareness) generate the same outcome. Hence linguistic ambiguity, even when coupled with unawareness, is not sufficient justification for the cp doctrine, if the unawareness is of the slip-the-mind variety. In the genuinely-clueless case, on the other hand, we show that a case can be made in favor of the CP doctrine.

4.3.1. The Slip-the-Mind Case

When agent 2 reads a contract that mentions objects that he was previosuly unaware of, and if he was unaware of them simply because they slipped his mind, he will become aware of those objects after he reads the contract. Suppose the contract is (g,h,ϕ). Let S be the set of objects mentioned in the sentence ϕ; i.e., $S = sen(l_1(\phi)) = sen(l_2(\phi)) = sen(d(\phi))$ for both interpretive doctrines d. Before agent 2 reads the contract, his awareness function is $\mathscr{A}_2(w) = \{\iota\}$ for all w; after he reads the contract, his awareness function becomes $\mathscr{A}_2(w) = \{\iota\} \cup S$ for all w.

Recall that $E = (Z,\iota)$ is the event "agent 2 suffers an income shock". So

[13] See, for example, Filiz-Ozbay (2012), Ozbay (2007), and Tirole (2009).

the four events

$$E \wedge d(\phi), \quad E \wedge \neg d(\phi), \quad \neg E \wedge d(\phi), \quad \neg E \wedge \neg d(\phi),$$

that are relevant for agent 2's accept/reject decision all have the same *sense*, namely $\{\iota\} \cup S$. Since after reading the contract, $\mathscr{A}_2(w) = \{\iota\} \cup S$ for every w, agent 2 can form explicit probabilistic beliefs about these events. This allows him to calculate the expected utilities resulting from accepting and rejecting the contract.

A simple backward induction argument then suggests that the insurer, who is aware of every object throughout, will choose a $\phi \in \mathscr{L}$ such that $ref(d(\phi))$ best approximates Z, and internalizes the gains from trade by setting the insurance premium at the level that makes agent 2 indifferent between accepting and rejecting. As in the benchmark case, the insurer's ability to approximate an arbitrary Z is restricted by the contractual language, and the exact restrictions depend on the interpretive doctrine d. This is captured by the fact that both $\mathscr{R}_{cp}$ and $\mathscr{R}_{anti-cp}$ are in general strictly smaller than 2^W.

By Proposition 4, we know that $|\mathscr{R}_{anti-cp}| = |\mathscr{R}_{cp}|$, so neither doctrine systematically outperforms the other. Either $\mathscr{R}_{anti-cp} = \mathscr{R}_{cp}$ (in which case the choice of the interpretive doctrine is irrelevant), or $\mathscr{R}_{anti-cp} \setminus \mathscr{R}_{cp} \neq \varnothing$ (in which case one can readily construct an example where full insurance is possible only under the anti-cp doctrine).

4.3.2. The Genuinely-Clueless Case

To help understand the clueless case, consider the example of a pet insurance policy. Such policies typically list the various diseases that are covered by the policy.[14] The list contains diseases such as balanoposthitis, esophagitis, enteritis, enucleation, FIP, HGE, hemobartonella, histiocytoma, leptospirosis, neoplasia, nephrectomy, pneumothorax, pyothorax, rickettsial, tracheobronchitis .... Most insureds have no idea what these diseases are even *after* reading the insurance contract. This is exactly what we assume in the clueless case, where agent 2's awareness function is the same before as after reading the contract; i.e., $\mathscr{A}_2(w) = \{\iota\}$ for all w.

A knee-jerk intuition may suggest that no contract with a positive premium will be accepted by agent 2, because he cannot fully understand it. "If

[14] See, for example, the policies offered at www.petinsurance.com.

I am offered a contract that reads $(\$10, \$100,$ "Barney catches disease xxx"),"
the knee-jerk intuition argues, "then the chances are that Barney will never
catch xxx, and the insurer will never need to pay me anything." We shall see
shortly that the knee-jerk intuition is half right but also half wrong. Under-
standing why it is half wrong is the key to understanding why the cp doctrine
is the superior interpretive doctrine.

Consider two different insurance policies, one covering balanoposthitis
but not tracheobronchitis, and the other covering tracheobronchitis but not
balanoposthitis. These two policies clearly differ, but the insured would not
be able to base his accept/reject decision on the basis of this difference if he
unaware of both diseases. Suppose he knows that some diseases are common
and expensive to treat, while others are rare and inexpensive to treat. If the
insured takes into account that the insurance policy is written by a rational
insurer, who in turn knows that the insured is unaware of either disease, then
a simple game-theoretic argument would enable the insured to figure out that
the disease covered in the actual contract he is offered must be the less ex-
pensive one. Note that agent 2's pessimism does not follow logically from
unawareness per se, but rather from the analysis of his opponent's strategic
behavior.

This informal argument suggests that we can analyze the clueless case by
representing it as an imperfect information game. Agent 1's actions are the
different contracts she can write. Agent 2 does not perfectly observe agent
1's action. But those actions are partitioned into different information sets
for agent 2. A contract that covers only balanoposthitis belongs to the same
information set as another contract that covers only tracheobronchitis (assum-
ing it has the same premium and payment as the first one), and both are in a
different information set from a third contract that covers both balanoposthitis
and tracheobronchitis, which in turn belongs to the same information set as
a fourth contract that covers leptospirosis and brucellosis, and so on. In any
(perfect Bayesian) equilibrium of such a game, agent 2 must hold pessimistic
beliefs with respect to any information set on the equilibrium path.

Let's illustrate this idea using a simple example, which also serves to
counter the knee-jerk intuition above.

In this simple example, l_1 is the same as l_2, so there is no linguistic am-
biguity and the choice of interpretive doctrine is irrelevant (we are merely
trying to demonstrate that some insurance is possible even under asymmet-
ric unawareness). So there is no need to distinguish predicates and proper-

ties. There are three states: $W = \{w_1, w_2, w_3\}$. Agent 2 suffers an income shock in states w_1 and w_2: $E = (\{w_1, w_2\}, \iota)$. There are infinitely many objects: $O = \{\iota, a, b, c, \ldots\}$, but $O_w = \hat{O} = \{\iota, a, b\}$ for all w. There is only one predicate/property: P, with $P(a) = (\{w_1, w_2\}, a)$, $P(b) = (w_1, b)$, and $P(x) = (\varnothing, x)$ for $x = c, d, \ldots$ As stated above, we assume that $\mathscr{I}_1(w) = \mathscr{I}_2(w) = W$, $\mathscr{A}_1(w) = O$, and $\mathscr{A}_2(w) = \{\iota\}$ for all w. The prior q puts equal probability on each state.

In this example, agent 2 explicitly knows that agent 1 is aware of some objects that he is unaware of; indeed, he explicitly knows that the number of such objects is exactly two (see the discussion following equation (8) above). He explicitly knows that there exists *something* that satisfies property P most of the time, although he is unaware of what it is. He also explicitly knows that there exists something else that satisfies property P less often, but at least whenever that *something* satisfies P he will also suffer an income shock. More importantly, he explicitly knows that there does not exist anything that never satisfies P. Thus when he sees a contract of the form $(g, 1, P(\cdot))$, where g satisfies

$$3U(1-g) \geq 2U(1) + U(0), \tag{9}$$

he will be willing to accept the contract even though he is unaware of the specific object mentioned in the contract. In equilibrium, the insurer will offer the contract $(g^*, 1, P(b))$ such that g^* satisfies (9) with equality.[15]

The above example is a counter-example to the knee-jerk intuition. Although it is natural to think of the set O as being very large,[16] $\hat{O}$ need not be, or at least agent 2 need not believe that it is. If agent 2 believes that there are not that many things that he is unaware of, he would be less worried about being tricked. The initial appeal of the knee-jerk intuition comes from an implicit assumption that $\hat{O}$ is big. We shall call this the rich-object assumption, and formalize it as follows. For any sentence $\phi \in \mathscr{L}$, the events $l_1(\phi)$, $l_2(\phi)$, $d_{cp}(\phi)$, and $d_{anti-cp}(\phi)$ all have the same (nonempty) sense, call it S. Suppose $S = \{a_1, \ldots, a_n\}$, and write ϕ as $\phi[a_1, \ldots, a_n]$ to make this explicit. From any

[15] It is important to understand why the insurer will not offer, for instance, the contract $(g^*, 1, P(c))$, even though such a contract will also be accepted by the insured. There is no real object that bears the name "c" that the insurer can point to to prove to the court that $P(c)$ does not obtain; given that the burden of proof in on the insurer to show that he does not have to payout, he will have to payout in every state.

[16] O is the set of hypothetical as well as real objects, and hence is limited only by our agents' imagination.

sentence $\phi[a_1,\ldots,a_n]$, and any n distinct objects $b_1,\ldots,b_n$, we can construct another sentence $\phi[b_1,\ldots,b_n]$ which is the same as $\phi[a_1,\ldots,a_n]$ with each a_j replaced by b_j. It is easy to verify that $\phi[b_1,\ldots,b_n]$ is also an element of $\mathscr{L}$.

Assumption 5 (The Rich-Object Assumption)**.** *Let d denote the interpretive doctrine used by the court. For any sentence $\phi[a_1,\ldots,a_n] \in \mathscr{L}$, either $ref(d(\phi[a_1,\ldots,a_n])) = W$, or there exist n distinct objects, $b_1,\ldots,b_n$, such that*

1. *$b_1,\ldots,b_n \in \hat{O}$, and*
2. *$ref(d(\phi[b_1,\ldots,b_n])) = \varnothing$.*

Note that the Rich-Object Assumption is a joint assumption on $\hat{O}$ and the interpretive doctrine d: fixing $\mathscr{L}$, l_1, and l_2, $\hat{O}$ may satisfy the Rich-Object Assumption under one doctrine d but not under another. The importance of the Rich-Object Assumption is summarized by the following proposition, the first part of which formalizes the knee-jerk intuition.

Proposition 6. *Let d denote the interpretive doctrine used by the court.*

1. *If the Rich-Object Assumption holds, then in any perfect Bayesian equilibrium, agent 2 receives no insurance.*
2. *If the Rich-Object Assumption does not hold, then there exists nonempty $R \subseteq W$ such that, if agent 2 suffers an income shock exactly in states in R, then there exists a perfect Bayesian equilibrium where agent 1 offers a contract that fully insures agent 2, and agent 2 accepts it.*

Proof. 1. Suppose $(g,h,\phi[a_1,\ldots,a_n])$ is a contract that is both offered and accepted with positive probability in any equilibrium.
If $ref(d(\phi[a_1,\ldots,a_n])) = W$, then the fact that it is offered with positive probability in equilibrium implies that $h \leq g$, and hence agent 2 receives no insurance under this contract. Suppose $ref(d(\phi[a_1,\ldots,a_n])) \subsetneq W$. Then $(g,h,\phi[b_1,\ldots,b_n])$, where $\phi[b_1,\ldots,b_n]$ is as defined in the Rich-Object Assumption, will also be accepted with positive probability. However, by the Rich-Object Assumption, agent 1 can always prove that the event $d(\phi[b_1,\ldots,b_n])$ does not obtain and hence avoid paying the insurance premium h. The fact that the original contract is offered with

positive probability implies that agent 1 also never needs to pay the insurance premium under that contract. Hence agent 2 receives no insurance from it.

2. Let $\phi[a_1,\ldots,a_n]$ be a sentence that invalidates the Rich-Object Assumption. Let $(b_1^*,\ldots,b_n^*)$ be a solution of the following minimization problem:
$$\min_{\underbrace{b_1,\ldots,b_n}_{distinct}\in\hat{O}} q(ref(d(\phi[b_1,\ldots,b_n]))),$$
where the existence of a solution is guaranteed by the finiteness of W. Finally, define R to be $ref(d(\phi[b_1^*,\ldots,b_n^*]))$. By assumption, R is nonempty. Then, if agent 2 suffers an income shock exactly in states in R, contracts of the form $(g,1,\phi[b_1^*,\ldots,b_n^*])$ will fully insure agent 2. A simple argument then establishes the existence of a perfect Bayesian equilibrium where agent 1 offers this contract with the insurance premium g such that agent 2 is indifferent between accepting and rejecting, and agent 2 accepts the contract. The fact that $(b_1^*,\ldots,b_n^*)$ solves the above minimization problem implies that agent 1 cannot profitably deviate to other contracts within the equivalence class of $\{(g,1,\phi[b_1,\ldots,b_n]) \mid b_1,\ldots,b_n \text{ distinct}\}$. □

We can now formalize the benefit of the cp doctrine over the anti-cp doctrine: the cp doctrine minimizes the chance that the Rich-Object Assumption holds.

Proposition 7. *Whenever the Rich-Object Assumption holds under the cp doctrine, it will also hold under the anti-cp doctrine.*

Proof. It suffices to observe that, for any $\phi\in\mathscr{L}$, $ref(d_{anti-cp}(\phi))\subseteq ref(d_{cp}(\phi))$. □

The converse of Proposition 7 is not true, as illustrated by the following simple example.

EXAMPLE: In this example, there are two states, $W=\{w_1,w_2\}$, two contractible objects, a and b, and one predicate, P. The two interpretations of P are as follows:

$$\begin{array}{lcllcl} l_1(P(a)) & = & (w_1,a), & l_1(P(b)) & = & (w_2,b),\\ l_2(P(a)) & = & (\varnothing,a), & l_2(P(b)) & = & (W,b). \end{array}$$

Suppose $Z = \{w_1\}$. Then, under the cp-doctrine, agent 1 can offer a contract $(g,h,P(a))$, with appropriately g and h, and fully insures agent 2. (Full insurance is achieved because $d_{cp}(P(a)) = (w_1,a)$.) Even when agent 1 anticipates that agent 2 will accept both contracts $(g,h,P(a))$ and $(g,h,P(b))$, as he cannot distinguish the two, she will have no incentive to deviate to offering contract $(g,h,P(b))$, because $d_{cp}(P(a)) = (W,a)$. The same is not true under the anti-cp doctrine. Indeed, it is a mechanical exercise to check that the Rich-Object Assumption is satisfied under the anti-cp doctrine. For example, if agent 1 anticipates that agent 2 will accept the contract $(g,h,P(b))$, she will deviate to contract $(g,h,P(a))$, because $d_{anti-cp}(P(b)) = (w_2,b)$, while $d_{anti-cp}(P(a)) = (\varnothing,b)$. Similarly, if agent 1 anticipates that agent 2 will accept the contract $(g,h,P(b) \wedge \neg P(a))$, she will deviate to contract $(g,h,P(a) \wedge \neg P(b))$, because $d_{anti-cp}(P(b) \wedge \neg P(a)) = (w_2,\{a,b\})$, while $d_{anti-cp}(P(a) \wedge \neg P(b)) = (\varnothing,\{a,b\})$.

4.4. Discussion

1. In the above analysis, we compared the cp doctrine only with the anti-cp doctrine. Ideally, we would like to define a general class of interpretive doctrines, and establish the cp doctrine as the optimal one among them. This is a task for future research. Here, we briefly remark on what care one should take when pursuing this problem. Consider a constant "doctrine", d, that maps any contractual sentence to the same event with a non-empty reference, say R. Under such a "doctrine", the rich-object assumption will never hold; and, with luck, Z may happen to be the same of R, making perfect insurance possible. Should d be in the feasible set of the optimal doctrine design problem? One may argue not, because d is insensitive to society's interpretations of contractual language, and hence is hardly a legal *interpretive* doctrine. But then what is the appropriate definition for legal interpretive doctrines? This is a question that a full-blown optimal doctrine design exercise needs to address first.
   A reasonable approach would be to define a legal interpretive doctrine as any function d such that $d(\phi) \in \{l_1(\phi), l_2(\phi)\}$ for every $\phi \in \mathscr{L}$. Under this definition, Proposition 7 can be strengthened as follows.

   Proposition 8. *Whenever the Rich-Object Assumption holds under the cp doctrine, it will also hold under any legal interpretive doctrine.*

The proof is the same as that of Proposition 7.

2. Our rationale for the cp doctrine actually does not depend on the assumption that the drafter of the contract has strictly richer awareness than the other party. For example, our argument continues to go through even if agent 2 is also aware of an array of (real) objects that agent 1 is not aware of. Those objects will play no role in the analysis, because the drafter, by definition, cannot write any sentence that makes reference to objects that she is unaware of. Additionally, suppose that there is an array of (real) objects that both agents 1 and 2 are aware of. The rationale behind the cp doctrine seems intuitive enough that it should be robust with respect to this complication as well, although the statements of the Rich-Object Assumption and of Proposition 6 would not be as clean.

3. Our analysis of the slip-the-mind case may seem surprising to the reader, especially in light of the recent literature where various authors have obtained interesting results in insurance contract design when the insured lacks full awareness. Let's point out an implicit assumption that differentiates our work from the rest. We assume that, after agent 2 reads a contract that reminds him of some objects that had previously slipped his mind, he continues to assign the same probability to the event of a negative income shock as before. If this assumption seems implausible, recall that in our framework it is possible for an agent to (explicitly) believe that *something* has slipped his mind, even though he is not aware of anything that has; hence he is not surprised when he later on comes across an example of such a thing. An agent's awareness and his (implicit) beliefs are logically distinct. While one could also tell stories where there is some link between the two, our present aim is to consider what difficulties are imposed on contracting parties by lack of awareness alone. To this end, we work with a model that captures this issue but isolates it from all others. We recognize that a fully-fledged theory of insurance contracts would need to address more systematically the question of how an agent's knowledge, probabilistic beliefs, and awareness change when he is exposed to new information that makes reference to objects that he was unaware of earlier. Developing models that do just this is a priority for our future research.

5. SPECULATIVE TRADE

In this section, we use the OBU structures to study the possibility of speculative trade under unawareness.[17] It is well known that, in classical state-space models with a common prior, common knowledge of strict willingness to trade is impossible when agents are non-delusional (i.e., if they never hold false belief [18]); on the other hand, when agents are delusional, speculative trade may occur. This result remains true when there is unawareness. Here we present two new results that we believe will be of some interest: either if everyone is *living in denial* (i.e., believes, perhaps incorrectly, that there is nothing they are unaware of), or if everyone is *living in paranoia* (i.e., believes, perhaps incorrectly, that there is something that they are unaware of), common knowledge of strict willingness to trade is still impossible, notwithstanding the fact that the agents may be delusional. The proof of this result makes use of an auxiliary theorem which is of interest on its own. The auxiliary theorem states that speculative trade is impossible as long as agents are *terminally partitional*, and hence generalizes the c6lassical no-trade theorem even in standard state-space models.[19]

5.1. Review of the Classical No-Trade Theorem

An OBU structure with common prior is given by $\langle W, O, \{O_w\}, \{\mathscr{I}_i\}, \{\mathscr{A}_i\}, q\rangle$, where W is finite (see Section 3). For the remainder of this section we assume that the information functions $\mathscr{I}_i$ satisfy *belief consistency*, i.e. for all $w \in W$ and all i, $\mathscr{I}_i(w) \neq \varnothing$. Belief consistency guarantees that conditional expectations are well defined. Given any OBU structure with common prior, we shall

[17] Heifetz et al. (2013) also study the possibility of speculative trade under unawareness, in a rather different framework from our own. They do not study situations where agents are living in denial or in paranoia.

[18] So far, we have been talking about what an agent knows and does not know, and interpreting L_i and K_i as *knowledge* operators. But these operators can also be interpreted as representing what an agent believes. Typically, it is assumed that one of the differences between knowledge and belief is that while truth is a necessary condition for knowledge, one may believe in something that is false. Since the main aim of this section is to analyze the implications of various assumptions about what is true, it may be clearer and more appropriate to talk about belief in this section, and be very explicit about truth/falsehood.

[19] Geanakoplos (1989) provides other generalizations of the classical no-trade theorem. The five conditions studied there (nondelusion, knowing that you know, nestedness, balancedness, and positively balancedness) neither imply nor are implied by terminal partitionality.

call the corresponding pair $\langle W, \{\mathscr{I}_i\}\rangle$ its *Kripke frame* (after the logician Saul Kripke).

With two additional restrictions on the information functions, Kripke frames form the basis of the standard model of information used in the economics literature:

- Non-delusion: for all $w \in W$ and all i, $w \in \mathscr{I}_i(w)$.
- Stationarity: for all $w, w' \in W$ and all i, if $w' \in \mathscr{I}_i(w)$ then $\mathscr{I}_i(w) = \mathscr{I}_i(w')$.

We refer to these two assumptions jointly as *partitionality*, since together they imply that $\mathscr{I}_i$ defines a partition on W. A Kripke frame that satisfies non-delusion and stationarity is often referred to as an *Aumann structure* or *information partition model*. Intuitively, non-delusion implies that if an agent (implicitly) believes a fact, then that fact is true; stationarity implies that agents believe that they believe what they actually do believe (positive introspection) and also believe that they don't believe what they actually don't believe (negative introspection).

Let $v : W \to \mathbb{R}^I$ be a function that satisfies $\sum_i v_i(w) = 0$ for every state w. The function v can be thought of as a trade contract that specifies the net monetary transfer to each agent in each state. Let F_i^v denote the event with empty sense (i.e., $sen(F_i^v) = \varnothing$) and with reference equal to the subset of worlds in which agent i's conditional expection of v is strictly positive:

$$ref(F_i^v) = \left\{ w \,\middle|\, \frac{\sum_{w' \in \mathscr{I}_i(w)} q(w') v_i(w')}{\sum_{w' \in \mathscr{I}_i(w)} q(w')} > 0 \right\}.$$

F_i^v can be interpreted as the event that agent i has strict willingness to trade. Let F^v be the conjunction of F_i^v's for every i (i.e., $F^v = \wedge_i F_i^v$), so that F^v is the event that every agent has strict willingness to trade. Let $\mathsf{K}^n F^v$ be recursively defined as $\wedge_i \mathsf{K}_i \mathsf{K}^{n-1} F^v$, with $\mathsf{K}^0 F^v = F^v$. Finally, define

$$\mathsf{CK} F^v := \wedge_{n \geq 1} \mathsf{K}^n F^v.$$

Clearly, $\mathsf{CK} F^v$ is the event that it is a common belief that every agent has strict willingness to trade. The *no-trade* happens if $ref(\mathsf{CK} F^v) = \varnothing$ for every trade contract v. On the other hand, if $w \in \mathsf{CK} F^v$ for some v and w, then *speculative trade* occurs.

The following result is a straightforward translation of the classical no-trade theorem to our setting. See, for example, Samet (1998) for a proof.

Proposition 9. *Take any OBU structure with common prior. If it satisfies non-delusional and stationarity (i.e., if it is partitional), then the no-trade result obtains.*

It is also well known that stationarity alone, without non-delusion, does not suffice to guarantee the no-trade result, nor does non-delusion alone without stationarity. In the next subsection, we prove a stronger version of the classical No-Trade Theorem, which says that the no-trade result still obtains when partitionality is weakened to a condition we call *terminal partitionality*.

5.2. Terminal Partitionality

Given any OBU structure, let $\langle W, \{\mathscr{I}_i\}\rangle$ be its Kripke frame. We first generalize the notion of partitionality to subspaces of W: $W' \subseteq W$ is partitional if for all $w, w' \in W'$, $\mathscr{I}_i(w) \subseteq W'$ for all i, and also non-delusion and stationarity are satisfied. Next, for every subspace $W' \subseteq W$, define

$$D(W') = \{w \in W \mid \mathscr{I}_i(w) \subseteq W' \text{ for some agent } i\}.$$

$D(W')$ is the collection of worlds in which at least one agent considers only worlds in W' to be possible. We say that an OBU structure (and its Kripke frame) satisfies *terminal partitionality* if there is a non-empty partitional subspace $W' \subseteq W$ such that $\cup_{n\geq 0} D^n(W') = W$, where $D^n(W')$ is defined recursively as $D(D^{n-1}(W'))$, and $D^0(W') = W'$.

Note that terminal partitionaity is a strictly weaker condition than partitionality. It says that there is a subset of states where agents satisfy non-delusion and stationarity (i.e. where everything they believe is true and they have access to their own beliefs), and in every other state, some agent either believes that everyone satisfies non-delusion and stationarity, or believes that someone believes that everyone satisfies non-delusion and stationarity, or believes that someone believes that someone believes that ....

The next proposition says that the condition of partitionality in the classical no-trade theorem can be replaced by terminal partitionality.

Proposition 10. *Take any OBU structure with common prior. If it is terminally partitional, then the no-trade result obtains.*

Proof. Let $\langle W, \{\mathscr{I}_i\}\rangle$ be the corresponding Kripke frame, and let W' be a partitional subspace such that $\cup_{n\geq 0} D^n(W') = W$. Such a partitional subspace

exists by assumption. We prove by induction that

$$ref(\mathsf{CK}F^v) \cap D^n(W') = \varnothing \tag{10}$$

for every n, which implies that $ref(\mathsf{CK}F^v) = ref(\mathsf{CK}F^v) \cap W = \varnothing$, completing the proof. For $n = 0$, this follows from Proposition 9 (applied to the sub-structure with state space W').

For the inductive step, suppose equation (10) has been proved up to n; we prove it for $n+1$. Consider any world $w \in D^{n+1}(W')$; i.e., any world w such that $\mathscr{I}_i(w) \subseteq D^n(W')$ for some agent i. Suppose $w \in ref(\mathsf{CK}F^v)$. Then $w \in ref(\mathsf{K}_i\mathsf{K}^m F^v)$ for every $m \geq 1$, and hence $\mathscr{I}_i(w) \subseteq ref(\mathsf{K}^m F^v)$ for every $m \geq 1$. Therefore $\mathscr{I}_i(w) \subseteq ref(\mathsf{CK}F^v)$. But then $ref(\mathsf{CK}F^v) \cap D^n(W') \supseteq \mathscr{I}_i(w) \neq \varnothing$ yields a contradiction. So we have $ref(\mathsf{CK}F^v) \cap D^{n+1}(W') = \varnothing$, as required. □

5.3. Living in Denial and Living in Paranoia

Informally, we say that an agent is *living in denial* if she always believes that there is nothing she is unaware of (although there may be). Similarly, we say that she is *living in paranoia* if she always believes that there is something she is unaware of (although there may be none). Let's illustrate these two concepts with two examples before getting into the formality.

Example 1. *Consider an OBU structure with only one agent;* $W = \{w_1, w_2\}$*;* $O = \{o_1, o_2\}$*,* $O_{w_1} = \{o_1\}$*,* $O_{w_2} = \{o_1, o_2\}$*;* $\mathscr{A}(w_1) = \mathscr{A}(w_2) = \{o_1\}$*; and* $\mathscr{I}(w_1) = \mathscr{I}(w_2) = \{w_1\}$.

In this example, although the agent is aware of exactly the same object in both states (i.e., $\mathscr{A}(w_1) = \mathscr{A}(w_2)$), different things are true in these states. In particular, in w_1 there is nothing that the agent is unaware of, while in w_2 there is something that the agent is unaware of. Note that in both states, the agent considers only w_1 as possible. Therefore the agent is delusional in w_2: she believes that there is nothing she is unaware of when there actually is. In this example, the agent always believes that there is nothing she is unaware of (although there may be), and hence she is living in denial.

Example 2. *Consider an OBU structure which is the same as in Example 1, except for that the information function is now* $\mathscr{I}(w_1) = \mathscr{I}(w_2) = \{w_2\}$.

In this example, in both states w_1 and w_2, the agent considers only w_2 possible. Therefore the agent is delusional in world w_1: she believes that there is something she is unaware of when there actually is none. In this example, the agent always believes that there is something she is unaware of (although there may be none), and hence she is living in paranoia.

Of course there is no reason why agents who are living in denial could not coexist with agents who are living in paranoia. An interesting task for future research is to study strategic interaction among these different kinds of agents. For now, however, we focus on cases where everyone is living in denial, or where everyone is living in paranoia.

Note that an agent who is living in denial may be delusional, and the classical no-trade theorem (Proposition 9) does not rule out the possibility of speculative trade. But living in denial, when it gives rise to delusion, results in a very specific form of delusion. In fact, we show that if this is the only form of delusion suffered by the agents, then speculative trade is still impossible. A similar result holds for the case where everyone is living in paranoia.

Definition 11. *An OBU structure satisfies WLID (weak living-in-denial) if, for every state w and agent i,*

1. $\mathscr{A}_i(w) \subseteq O_w$;
2. $\mathscr{A}_i(w') = O_{w'}$ *for every* $w' \in \mathscr{I}_i(w)$; *and*
3. $\mathscr{A}_i(w) = O_w$ *implies* $w \in \mathscr{I}_i(w)$ *and* $\mathscr{I}_i(w') = \mathscr{I}_i(w)$ *for* $w' \in \mathscr{I}_i(w)$.

The second part of the definition says that agent i considers possible only states in which she is aware of everything, and so she believes (correctly or incorrectly) that there is nothing she is unaware of. The third part says that if this belief turns out to be correct in a given state, then she has no false beliefs in that state and has access to her own beliefs.

Definition 12. *An OBU structure satisfies WLIP (weak living-in-paranoia) if, for every state w and agent i,*

1. $\mathscr{A}_i(w) \subseteq O_w$;
2. $\mathscr{A}_i(w') \subsetneq O_{w'}$ *for every* $w' \in \mathscr{I}_i(w)$; *and*
3. $\mathscr{A}_i(w) \subsetneq O_w$ *implies* $w \in \mathscr{I}_i(w)$ *and* $\mathscr{I}_i(w') = \mathscr{I}_i(w)$ *for* $w' \in \mathscr{I}_i(w)$.

WLIP is the opposite of WLID in some sense: every agent believes (correctly or incorrectly) that there is something she is unaware of; and if she turns out to be correct about this, she is correct on every other matter and also has access to her own beliefs.

Both WLID and WLIP are "weak" conditions in the sense that even a partitional OBU structure can satisfy WLID or WLIP (although it cannot satisfy both simultaneously).

Before we state our main results, we need one more definition. We say that an OBU structure satisfies *LA-introspection* if, for every state w and every agent i, $w' \in \mathscr{I}_i(w)$ implies $\mathscr{A}_i(w') = \mathscr{A}_i(w)$. LA-introspection is characterized by Board & Chung (2021)'s axioms **LA1** and **LA2**, which jointly say that every agent has correct beliefs about what she is aware of (see Board & Chung (2021) for more details).

Proposition 13. *Consider an OBU structure with common prior, and suppose it satisfies WLID and LA-introspection. Then it also satisfies terminal partitionality.*

Proof. For any two worlds, w and w', we say that *w points to w'* if there is an agent i such that $w \notin \mathscr{I}_i(w)$ and $w' \in \mathscr{I}_i(w)$.

Suppose w points to w'. Then $w \notin \mathscr{I}_i(w)$ for some agent i. By WLID, LA-introspection, and WLID again, we have

$$O_{w'} = \mathscr{A}_i(w') = \mathscr{A}_i(w) \subsetneq O_w \tag{11}$$

for some agent i. Therefore a world can only point to other worlds that have strictly smaller sets of real objects. Then, by finiteness of W, there exist worlds that do not point to any other worlds. Let W' be the collection of these worlds.

If w belongs to W', then $w \in \mathscr{I}_i(w)$ for any agent i. Furthermore, for any agent i, by the second and the third parts of WLID respectively, we have $\mathscr{A}_i(w) = O_w$ and hence $\mathscr{I}_i(w') = \mathscr{I}_i(w)$ for any $w' \in \mathscr{I}_i(w)$. But this means $w' \in \mathscr{I}_i(w)$ implies $w' \in \mathscr{I}_i(w')$, and hence w' also does not point to any other worlds. Therefore W' is a partitional subspace.

If $W \neq W'$, then by finiteness of $W \setminus W'$, and by the observation that a world can only point to worlds that have strictly smaller sets of real objects, there must exist worlds in $W \setminus W'$ that do not point to any other worlds in $W \setminus W'$. Let W'' be the collection of these worlds. It is easy to verify that

$D(W') = W'' \cup W' \supsetneq W'$. Repeating this argument, one can show that if $W \neq D^n(W')$, then $D^{n+1}(W')$ is a strict superset of $D^n(W')$. Therefore, by finiteness again, $W = \cup_{n \geq 0} D^n(W')$. □

Proposition 14. *Consider an OBU structure with common prior, and suppose it satisfies WLIP and LA-introspection. Then it also satisfies terminal partitionality.*

Proof. The proof is similar to that of Proposition 13, except for equation (11). Suppose w points to w'. Then $w \notin \mathscr{I}_i(w)$ for some agent i. By WLIP, LA-introspection, and WLIP again, we have

$$O_{w'} \supsetneq \mathscr{A}_i(w') = \mathscr{A}_i(w) = O_w$$

for some agent i. Therefore a world can only point to other worlds that have strictly larger sets of real objects. The rest of the proof now follows the same arguments as in that of Proposition 13. □

Corollary 15. *Consider a regular OBU structure with common prior, and suppose it satisfies LA-introspection. If it satisfies either WLID or WLIP, then the no-trade result obtains.*

Proof. This follows from Propositions 10, 13 and 14. □

6. CONCLUSION

As we discussed in the introduction, there is large gap in the literature on unawareness between the more applied studies that appeal to unawareness to motivate the assumptions underlying their models, and the foundational studies that often pay little attention to the real-world applications. In this and our companion papers, we have attempted to bridge this gap. In particular, we have shown that the key assumption in the applied literature, namely that agents are "unaware, but know that they are unaware", can be captured in a rational-agent framework; furthermore, this assumption is perfectly consistent with the DLR axioms that much of the foundational literature tries to accommodate.

Although the OBU structures described above derive an agent's unawareness of propositions from her unawareness of the objects mentioned in those propositions, one can envisage an extension where unawareness of properties

is also modeled. A property-unawareness function could work (roughly) as follows: if an agent is unaware of a given property, then she would be unaware of any event containing one state but not another, where the two states could only be distinguished by whether or not various objects satisfied that property. Combining such a property-unawareness function with the object-unawareness function analyzed above would allow us to separate two kinds of unawareness: and agent could be unaware that "Yao Ming is tall" either because she has no idea who Yao Ming is or because she does not understand the concept of height.

In additional to providing foundations for a model of unawareness, in the form of OBU structures, we have also presented two applications: the first examines the legal interpretive doctrine contra proferentem, while the second extends the classical no trade theorem to cover cases where agents are mistaken in a particular away (they live in denial or in paranoia). These applications, we hope, will convince the reader that it is straightforward to use OBU structures in applied work. We also believe that the results of these applications are valuable in their own right.

Before we finish, we would also like to mention a recent experimental paper that provides evidence suggesting that agents may be unsure whether they are aware of everything or not. Blume & Gneezy (2010) have their subjects play a game with each other. There is a less-obvious strategy that guarantees a win, and a more-obvious strategy that results in a win half the time. Even though a win paid out \$10, some subjects rejected an outside option of \$6 and then played the more-obvious strategy, for an expected payout of \$5. Presumably these subjects were not aware of the less-obvious strategy. Why then did they reject the outside option? Blume & Gneezy (2010) suggest that this is because they believed such a strategy existed, and hoped to figure it out after rejecting the outside option but before playing the game. In our language, we would say that these agents believed there was something they were unaware of, and hoped to become aware of it in the future.

Appendix A: Model-Theoretic Description of OBU Structures

For the sake of transparency, and to aid interpretation, we now show how OBU structures assign truth conditions for a formal language, a version of

first-order modal logic.[20] We start with a set of (unary) predicates, $P, Q, R, \ldots$, and an (infinite) set of variables, $x, y, z, \ldots$. Together with set of objects, O, this generates a set Φ of atomic formulas, $P(\mathsf{a}), P(x), Q(\mathsf{a}), Q(x), \ldots$, where each predicate takes as its argument a single object or variable. Let $\mathscr{F}$ be the smallest set of formulas that satisfies the following conditions:

- if $\phi \in \Phi$, then $\phi \in \mathscr{F}$;
- if $\phi, \psi \in \mathscr{F}$, then $\neg\phi \in \mathscr{F}$ and $\phi \wedge \psi \in \mathscr{F}$;
- if $\phi \in \mathscr{F}$ and $x \in X$, then $\forall x\phi \in \mathscr{F}$;
- if $\phi \in \mathscr{F}$, then $\mathtt{L}_i\phi \in \mathscr{F}$ and $\mathtt{A}_i\alpha \in \mathscr{F}$ and $\mathtt{K}_i\alpha \in \mathscr{F}$ for each agent i.

Formulas should be read in the obvious way; for instance, $\forall x \mathtt{A}_i P(x)$ is to be read as "for every x, agent i is aware that x possesses property P." Notice, however, that it is hard to make sense of certain formulas: consider $P(x)$ as opposed to $P(a)$ or $\forall x P(x)$. Although it may be reasonable to claim that a specific object, a, is P, or that every x is P, the claim that x is P seems empty unless we specify which object variable x stands for. In general, we say that a variable x is free in a formula if it does fall under the scope of a quantifier $\forall x$, and define our language $\mathscr{L}$ to be the set of all formulas containing no free variables.[21] We use OBU structures to provide truth conditions only for formulas in $\mathscr{L}$, and not for formulas such as $P(x)$ that contain free variables.

Take an OBU structure $M = \langle W, O, \{O_w\}, \{\mathscr{I}_i\}, \{\mathscr{A}_i\}\rangle$, and augment it with an *assignment* $\pi(w)(P) \subseteq O$ of objects to every predicate at every state (intuitively, $\pi(w)(P)$ is the set of objects that satisfy predicate P). If a formula $\phi \in \mathscr{L}$ is true at state w of OBU structure M under assignment π, we write $(M, w, \pi) \vDash P(\mathsf{a})$; $\vDash$ is defined inductively as follows:

[20] Board & Chung (2021) provide the (model-theoretic) soundness and complete axiomatization.

[21] More formally, we define inductively what it is for a variable to be free in $\phi \in \mathscr{F}$:

- if ϕ is an atomic formula of the form $P(x)$ where x is a variable, then x is free in ϕ;
- x is free in $\neg\phi$, $\mathtt{K}_i\phi$, $\mathtt{A}_i\phi$, and $\mathtt{L}_i\phi$ iff x is free in ϕ;
- x is free in $\phi \wedge \psi$ iff x is free in ϕ or ψ;
- x is free in $\forall y\alpha$ iff x is free in ϕ and x is different from y.

$(M,w,\pi) \vDash P(\mathsf{a})$ iff $\mathsf{a} \in \pi(w)(P)$;

$(M,w,\pi) \vDash \neg\phi$ iff $(M,w,\pi) \nvDash \phi$;

$(M,w,\pi) \vDash \phi \wedge \psi$ iff $(M,w,\pi) \vDash \phi$ and $(M,w,\pi) \vDash \psi$;

$(M,w,\pi) \vDash \forall x\phi$ iff $(M,w,\pi) \vDash \phi[\mathsf{a}\backslash x]$ for every $a \in O_w$ (where $\phi[\mathsf{a}\backslash x]$ is ϕ with all free occurrences of x replaced with a);

$(M,w,\pi) \vDash \mathtt{A}_i\phi$ iff $\mathsf{a} \in \mathscr{A}_i(w)$ for every object a in ϕ;

$(M,w,\pi) \vDash \mathtt{L}_i\phi$ iff $(M,w',\pi) \vDash \phi$ for all $w' \in \mathscr{I}_i(w)$;

$(M,w,\pi) \vDash \mathtt{K}_i\phi$ iff $(M,w,\pi) \vDash \mathtt{A}_i\phi$ and $(M,w,\pi) \vDash \mathtt{L}_i\phi$.

Notice that there is a close connection between sentences of $\mathscr{L}$ and OBU events: for any given $\phi \in \mathscr{L}$, the reference of the corresponding OBU event is given by the set of states at which ϕ is true, while the sense is simply the set of objects in ϕ.

Appendix B: Proofs

Proof of Proposition 1. 1. Straightforward.

2. Take some A'_i which satisfies A1–A4, and define $\mathscr{A}_i$ as follows: $a \in \mathscr{A}_i(w)$ iff $w \in ref\mathsf{A}'_i(W,a)$. We need to show that $\mathsf{A}'_i(R,S) = \mathsf{A}_i(R,S)$. We consider two cases:
Case 1: $S \neq \varnothing$. Then

$$\begin{aligned}
\mathsf{A}'_i(R,S) &= \mathsf{A}'_i(W,S) \text{ (by A2)} \\
&= \wedge_{a\in S}\mathsf{A}'_i(W,a) \text{ (by A1)} \\
&= \wedge_{a\in S}\left(\{w \mid x \in \mathscr{A}_i(w)\},a\right) \text{ (by A4 and the definition of } \mathscr{A}_i) \\
&= \left(\{w \mid S \subseteq \mathscr{A}_i(w)\},S\right) \text{ (definition of } \wedge) \\
&= \mathsf{A}_i(R,S), \text{ as required.}
\end{aligned}$$

Case 2: $S = \varnothing$. Then

$$\begin{aligned}
\mathsf{A}'_i(R,\varnothing) &= (W,\varnothing) \text{ (by A3)} \\
&= \left(\{w \in W \mid \varnothing \subseteq \mathscr{A}_i(w)\},\varnothing\right) \\
&= \mathsf{A}_i(R,\varnothing), \text{ as required.}
\end{aligned}$$

□

Proof of Proposition 2. 1. Straightforward.

2. Take some L_i' which satisfies L1–L4, and define $\mathscr{I}_i$ as follows:

$$\mathscr{I}_i(w) = \left\{w' \mid w \in ref\left(\neg \mathsf{L}_i' \neg\left(w', O\right)\right)\right\}.$$

Note that, by L4,

$$\left\{w' \mid w \in ref\left(\neg \mathsf{L}_i' \neg\left(w', O\right)\right)\right\} = \left\{w' \mid w \in ref\left(\neg \mathsf{L}_i' \neg\left(w', S\right)\right)\right\}$$

for all $S \subseteq O$, so $w' \in \mathscr{I}_i(w)$ iff $w \in ref\left(\neg \mathsf{L}_i' \neg\left(w', S\right)\right)$, and hence

$$w' \notin \mathscr{I}_i(w) \text{ iff } w \in ref\left(\mathsf{L}_i' \neg\left(w', S\right)\right). \qquad (*)$$

We need to show that $\mathsf{L}_i'(R,S) = \mathsf{L}_i(R,S)$. We consider two cases:
Case 1: $R \neq W$. Then

$$\begin{aligned}
\mathsf{L}_i'(R,S) &= \mathsf{L}_i'\left(\cap_{w \notin R} W \setminus \{w\}, S\right) \\
&= \wedge_{w \notin R} \mathsf{L}_i'\left(W \setminus \{w\}, S\right) \text{ (by L2)} \\
&= \wedge_{w \notin R} \mathsf{L}_i' \neg\left(w, S\right) \text{ (definition of } \neg) \\
&= \wedge_{w \notin R}\left(\left\{w' \mid w \notin \mathscr{I}_i\left(w'\right)\right\}, S\right) \text{ (by } (*) \text{ and L3)} \\
&= \left(\cap_{w \notin R}\left\{w' \mid w_1 \notin \mathscr{I}_i\left(w'\right)\right\}, S\right) \text{ (definition of } \wedge) \\
&= \left(\left\{w' \mid \mathscr{I}_i\left(w'\right) \subseteq R\right\}, S\right) \\
&= \mathsf{L}_i(R,S), \text{ as required.}
\end{aligned}$$

Case 2: $R = W$. Then $\mathsf{L}_i'(W,O) = (W,O)$ (by L1), so $\mathsf{L}_i'(W,S) = (W,S)$ (by L4). And $\mathsf{L}_i(W,S) = (\{w \mid \mathscr{I}_i(w) \subseteq W\}, S) = (W,S)$.

□

Proof of Proposition 3. 1. Straightforward.

2. Take some All' which satisfies All1–All4. For any $w \in W$ and $a \in O$, construct the property p_{wa} such that:

$$p_{wa}(b) = \begin{cases} (W, b) & \text{if } b \neq a \\ (W \setminus \{w\}, b) & \text{if } b = a \end{cases}.$$

Observe for later use that, by All2, $W \setminus \{w\} \subseteq ref(\mathsf{All}'\, p_{wa})$, and hence, for any $R \subseteq W$,

$$\cap_{w \notin R} ref(\mathsf{All}'\, p_{wa}) = \{w | w \in ref(\mathsf{All}'\, p_{wa})\} \cup R. \tag{12}$$

We define $\{O_w\}_{w \in W}$ using these p_{wa}'s as follows:

$$O_w = \left\{ a \mid w \notin ref\left(\mathsf{All}'\, p_{wa}\right) \right\}.$$

These O_w's define the property *re*:

$$R_a^{re} = \{w \mid w \notin ref(\mathsf{All}'\, p_{wa})\}.$$

This property *re*, of course, in turn defines the operator All. We need to show that $\mathsf{All}' = \mathsf{All}$. Take an arbitrary property $\tilde{p}$. From All4, we have $sen(\mathsf{All}'\, \tilde{p}) = S^{\tilde{p}}$; and $sen(\mathsf{All}\, \tilde{p}) = S^{\tilde{p}}$ from the definition of All. It remains to show that $ref(\mathsf{All}'\, \tilde{p}) = ref(\mathsf{All}\, \tilde{p})$.
From $\tilde{p}$, construct another property $\hat{p}$ as follows:

$$\hat{p} := \wedge_{a \in O} \wedge_{w \notin R^{\tilde{p}} a}\ p_{wa}.$$

We claim that $R_b^{\hat{p}} = R_b^{\tilde{p}}$ for every $b \in O$, and hence by All3, we have $ref(\mathsf{All}'\, \hat{p}) = ref(\mathsf{All}'\, \tilde{p})$. To prove this claim, notice that, for any $b \in O$,

$$\begin{aligned} R_b^{\hat{p}} &= \cap_{a \in O} \cap_{w \notin R_a^{\tilde{p}}} R_b^{p_{wa}} \\ &= (\cap_{\substack{a \neq b \\ w \notin R_a^{\tilde{p}}}} R_b^{p_{wa}}) \cap (\cap_{\substack{a = b \\ w \notin R_a^{\tilde{p}}}} R_b^{p_{wa}}) \\ &= (\cap_{\substack{a \neq b \\ w \notin R_a^{\tilde{p}}}} W) \cap (\cap_{w \notin R_b^{\tilde{p}}} W \setminus \{w\}) \\ &= R_b^{\tilde{p}}, \text{ as required.} \end{aligned}$$

Therefore, it suffices to prove that $ref(\mathsf{All}'\,\hat{p}) = ref(\mathsf{All}\,\tilde{p})$. By All1, we have

$$\begin{aligned}
ref(\mathsf{All}'\,\hat{p}) &= \cap_{a\in O}\cap_{w\notin R_a^{\tilde{p}}} ref(\mathsf{All}'\,p_{wa}) \\
&= \cap_{a\in O}\left(\{w \mid w \in ref(\mathsf{All}'\,p_{wa})\}\cup R_a^{\tilde{p}}\right) \quad \text{(by (12))} \\
&= \cap_{a\in O}\left(R_a^{\neg re}\cup R_a^{\tilde{p}}\right) \\
&= \cap_{a\in O} R_a^{\neg re\vee\tilde{p}} \\
&= \cap_{a\in O} R_a^{re\to\tilde{p}} \\
&= ref(\mathsf{All}\,\tilde{p}), \text{ as required.}
\end{aligned}$$

□

References

Abraham, K. S. (1996). A theory of insurance policy interpretation. *Michigan Law Review*, *95*(3), 531–569.

Blume, A., & Gneezy, U. (2010). Cognitive forward induction and coordination without common knowledge: An experimental study. *Games and Economic Behavior*, *68*(2), 488-511.

Board, O. J., & Chung, K.-S. (2021). Object-based unawareness: Axioms. *Journal of Mechanism and Institution Design*, *6*(6), 1–36.

Carolina Care Plan Incorporated v. McKenzie. (2007). *551 U.S. 1* (No. 06-1182 (R46-022)).

Chung, K.-S., & Fortnow, L. (2016). Loopholes. *The Economic Journal*, *126*(595), 1774-1797.

Dekel, E., Lipman, B. L., & Rustichini, A. (1998). Standard state-space models preclude unawareness. *Econometrica*, *66*(1), 159–173.

Fagin, R., & Halpern, J. Y. (1987). Belief, awareness, and limited reasoning. *Artificial Intelligence*, *34*(1), 39-76.

Filiz-Ozbay, E. (2012). Incorporating unawareness into contract theory. *Games and Economic Behavior*, *76*(1), 181-194.

Geanakoplos, J. (1989). Game Theory Without Partitions, and Applications to Speculation and Consensus. *Cowles Foundation Discussion Papers*, *914*.

Gul, F. (1998). A comment on Aumann's Bayesian view. *Econometrica*, *66*(4), 923–927.

Halpern, J. Y. (1999). Set-theoretic completeness for epistemic and conditional logic. *Annals of Mathematics and Artificial Intelligence*, *26*(1), 1-27.

Halpern, J. Y. (2001). Alternative semantics for unawareness. *Games and Economic Behavior*, *37*(2), 321-339.

Halpern, J. Y., & Rêgo, L. C. (2006). Reasoning about knowledge of unawareness. In *Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning* (pp. 6–13). AAAI Press.

Heifetz, A., Meier, M., & Schipper, B. C. (2006). Interactive unawareness. *Journal of Economic Theory*, *130*(1), 78-94.

Heifetz, A., Meier, M., & Schipper, B. C. (2013). Unawareness, beliefs, and speculative trade. *Games and Economic Behavior*, *77*(1), 100-121.

Li, J. (2009). Information structures with unawareness. *Journal of Economic Theory*, *144*(3), 977-993.

Modica, S., & Rustichini, A. (1994). Awareness and partitional information structures. *Theory and Decision*, *37*(1), 107-124.

Modica, S., & Rustichini, A. (1999). Unawareness and partitional information structures. *Games and Economic Behavior*, *27*(2), 265-298.

Morris, S. (1995). The common prior assumption in economic theory. *Economics and Philosophy*, *11*(2), 227–253. doi: 10.1017/S0266267100003382

Ozbay, E. Y. (2007). Unawareness and strategic announcements in games with uncertainty. In *Proceedings of the 11th Conference on Theoretical Aspects of Rationality and Knowledge* (p. 231–238). New York, NY, USA: Association for Computing Machinery.

Samet, D. (1998). Common priors and separation of convex sets. *Games and Economic Behavior*, *24*(1), 172-174.

Sillari, G. (2006). Models of awareness. In *Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision Theory (loft)* (pp. 209–218).

Tirole, J. (2009). Cognition and incomplete contracts. *American Economic Review*, *99*(1), 265-94.

Journal of Mechanism and Institution Design
ISSN: 2399-844X (Print), 2399-8458 (Online)
DOI:10.22574/jmid.2022.12.002

CENTRALIZED CLEARING MECHANISMS: A PROGRAMMING APPROACH

Péter Csóka

Corvinus University of Budapest and KRTK, Hungary
peter.csoka@uni-corvinus.hu

P. Jean-Jacques Herings

Tilburg University, Netherlands
P.J.J.Herings@tilburguniversity.nl

ABSTRACT

We consider financial networks where agents are linked to each other by financial contracts. A centralized clearing mechanism collects the initial endowments, the liabilities and the division rules of the agents and determines the payments to be made. A division rule specifies how the assets of the agents should be rationed. Since payments made depend on payments received, we are looking for solutions to a system of equations. The set of solutions is known to have a lattice structure, leading to the existence of a least and a greatest clearing payment matrix. Previous research has shown how decentralized clearing selects the least clearing payment matrix. We present a centralized approach towards clearing in order to select the greatest clearing payment matrix. To do so, we formulate the determination of the greatest clearing payment matrix as a programming problem. When agents use proportional division rules, this programming problem corresponds to a linear programming problem. We show that for other common division rules, it can be written as an integer linear programming problem.

Keywords: Systemic risk, bankruptcy rules, integer linear programming.

JEL Classification Numbers: C71, G10.

Both authors declare there are no conflicts of interest. We would like to thank Kolos Csaba Ágoston and participants of the Annual Conference of the Hungarian Society of Economics 2021 for helpful comments. Péter Csóka was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences. Any errors are our own.

1. INTRODUCTION

In this paper, we consider financial networks where agents are linked to each other by financial contracts. Like the seminal paper of Eisenberg & Noe (2001), a financial network consists of agents corresponding to the financial institutions, their initial endowments, and liabilities. An agent's initial endowment includes all the agent's tangible and intangible assets but excludes the claims and liabilities the agent has towards the other agents. For outstanding surveys of this literature, we refer the reader to Glasserman & Young (2016) and Jackson & Pernoud (2021).

In Eisenberg & Noe (2001), agents use proportional division rules to determine payments in case of bankruptcy, i.e., payments are proportional to liabilities. In practice, often priority principles are invoked, where a priority order determines the seniority of the liabilities. Given a permutation determining the rank of the claims, under the priority rule (see e.g.,Moulin (2000) and Chatterjee & Eyigungor (2015)) claimants are paid in a lexicographic order determined by the permutation. Other common division rules are the constrained equal awards rule and the constrained equal losses rule (for updated surveys, see Thomson (2013) and Thomson (2015)). Under the constrained equal awards division rule,[1] all claimants get the same amount, up to the value of their claim. The constrained equal losses division rule is its dual and imposes that all claimants face the same loss, up to the value of their claim. The choice of the division rule may also balance the trade-off between welfare maximization and payoff equalization (Gallice, 2019). Csóka & Herings (2018) note that on top of financial networks, default contagion can also occur in other applications (i.e. supply chains, international student exchange programs, servers processing job, time banks), where again other division rules may be in place. We will extend the Eisenberg & Noe (2001) framework and allow for general division rules.

In claims problems, there is a single, exogenously given, bankrupt agent and a division rule is used to determine the payments to the claimants. In financial networks, there can be multiple bankrupt agents. As an agent's asset value, and therefore payments made, depends on payments received, the actual payments are endogenously determined. Like the proportional rule for claims problems can be extended to financial networks (Eisenberg & Noe, 2001), it is possible to extend any division rule for claims problems to fi-

[1] For an axiomatization of its weighted version, see Flores-Szwagrzak (2015).

nancial networks (Groote Schaarsberg et al., 2018). The resulting payment matrix consists of first computing each agent's asset value as the sum of the initial endowments and the payments received and next making the payments in accordance with the given division rule.

Following Csóka & Herings (2021), a so-called clearing payment matrix satisfies the following three conditions. First, feasibility, which states that the payments are in accordance with the given division rules. Second, limited liability, which requires that the total payments made by an agent must never exceed the asset value of the agent. Third, priority of creditors, which expresses that default is only allowed if equity, i.e., asset value minus payments made, is equal to zero. Since payments made depend on payments received, the determination of a clearing payment matrix corresponds to the solution to a fixed point problem.

In this paper, we use the system of equations as introduced in Csóka & Herings (2021) to find a clearing payment matrix. The set of solutions to this system forms a complete lattice, which implies that there is a least and a greatest clearing payment matrix. Csóka & Herings (2018) show in a decentralized set-up that a large class of decentralized clearing processes converges to the least clearing payment matrix and Ketelaars & Borm (2021) derive an analogous result in a continuous set-up. In this paper, we therefore examine how a centralized approach can be used to select the greatest clearing payment matrix. More precisely, we present a programming problem whose unique solution is the greatest clearing payment matrix. The programming problem can be written as a linear programming problem when all agents use proportional division rules. For the other common division rules, we demonstrate how the programming problem reduces to an integer linear programming problem.

The rest of the paper is organized as follows. Section 2 defines financial networks and clearing payment matrices. Section 3 illustrates the possible multiplicity of clearing payment matrices and how multiplicity may vary with the division rules that are in place. Section 4 formulates the programming problems. Section 5 makes some concluding remarks.

2. FINANCIAL NETWORKS

A *financial network* is a quadruple $F = (N, z, L, d)$ with the following interpretation.

The finite set N consists of the *agents* in the financial network.

The vector $z \in \mathbb{R}^N_+$ represents the *endowments* of the agents, which are non-negative real numbers. The endowments of an agent include all the agent's tangible and intangible assets, but exclude the claims the agent has on the other agents.

The non-negative *liability matrix* $L \in \mathbb{R}^{N\times N}_+$ describes the mutual claims of the agents. Its entry L_{ij} is the liability of agent $i \in N$ towards agent $j \in N$ or, equivalently, the claim of agent j on agent i. It is allowed that simultaneously agent j has a claim on agent i and agent i has a claim on agent j, so $L_{ij} > 0$ and $L_{ji} > 0$ can both hold at the same time. Agent do not have claims on themselves, so we set $L_{ii} = 0$ for every $i \in N$.

The determination of the payments to be made by the agents takes place by *division rules* $d = (d^i)_{i\in N}$. The division rule $d^i : \mathbb{R}_+ \to \mathbb{R}^N_+$ of agent $i \in N$ describes which payments agent i makes to the agents in N as a function of agent's i estate E_i. Payments are non-negative, bounded above by the liabilities, and are such that the sum of the payments is equal to the minimum of the estate and the sum of the liabilities, so it holds that, for every $E_i \in \mathbb{R}_+$, for every $j \in N$, $d^i_j(E_i) \leq L_{ij}$, and $\sum_{j\in N} d^i_j(E_i) = \min\{E_i, \sum_{j\in N} L_{ij}\}$. Moreover, for every $j \in N$, d^i_j is required to be weakly increasing in E_i. It is well-known that the weak monotonicity of d^i implies that it is continuous, see for instance Thomson (2003). The estate of an agent in a financial network depends on the payments received on outstanding claims and is therefore determined endogenously.

Important examples of division rules are the proportional, priority, constrained equal awards, and the constrained equal losses division rules.

The division rule d^i of agent $i \in N$ is equal to the *proportional division rule* if, for every $E_i \in \mathbb{R}_+$, it assigns to claimant $j \in N$ the amount

$$d^i_j(E_i) = \begin{cases} 0, & \text{if } L_{ij} = 0, \\ \min\{\frac{L_{ij}}{\sum_{k\in N} L_{ik}} E_i, L_{ij}\}, & \text{otherwise.} \end{cases}$$

Under the proportional division rule, the estate is divided in a proportional way over the claimants, up to the value of those claims.

The division rule d^i of agent $i \in N$ is equal to a *priority division rule* if there exists a permutation $\pi : N \to \{1, \ldots, |N|\}$, determining the rank of the claims, such that, for every $E_i \in \mathbb{R}_+$,

$$d^i_j(E_i) = \max\{0, \min\{L_{ij}, E_i - \sum_{k\in N | \pi(k)<\pi(j)} L_{ik}\}\},$$

where $\{k \in N | \pi(k) < \pi(j)\}$ is the set of agents ranked before j according to π. Under a priority division rule, claims are paid sequentially to agents $\pi^{-1}(1)$, $\pi^{-1}(2), \ldots$ as long as the estate of agent i permits this.

We next define the constrained equal awards rule. Let $i \in N$. If $E_i > \sum_{j \in N} L_{ij}$, then define the award $\lambda_i = \max_{j \in N} L_{ij}$. Otherwise, define the award $\lambda_i \in [0, \max_{j \in N} L_{ij}]$ as the unique solution to

$$\sum_{j \in N} \min\{L_{ij}, \lambda_i\} = E_i.$$

The division rule d^i of agent i is equal to the *constrained equal awards division rule* if, for every $E_i \in \mathbb{R}_+$, it assigns to claimant $j \in N$ the amount

$$d^i_j(E_i) = \min\{L_{ij}, \lambda_i\}.$$

Under the constrained equal awards division rule, all claimants get the same amount, up to the value of their claim.

The constrained equal losses rule is the dual of the constrained equal awards rule. If $E_i > \sum_{j \in N} L_{ij}$, then define the loss $\mu_i = 0$. Otherwise, define the loss $\mu_i \in [0, \max_{j \in N} L_{ij}]$ as the unique solution to

$$\sum_{j \in N} \max\{L_{ij} - \mu_i, 0\} = E_i.$$

The division rule d^i of agent $i \in N$ is equal to the *constrained equal losses division rule* if, for every $E_i \in \mathbb{R}_+$, it assigns to claimant $j \in N$ the amount

$$d^i_j(E_i) = \max\{L_{ij} - \mu_i, 0\}.$$

Under the constrained equal losses division rule, all claimants face the same loss, up to the value of their claim.

The set of all matrices in $\mathbb{R}_+^{N \times N}$ with a zero diagonal is denoted by $\mathscr{M}$. The partial order $\leq$ on $\mathscr{M}$ is defined in the usual way: For $P, P' \in \mathscr{M}$ it holds that $P \leq P'$ if and only if $P_{ij} \leq P'_{ij}$ for all $(i, j) \in N \times N$. For $P \in \mathscr{M}$ and $i \in N$, let $P_i \in \mathbb{R}^N$ denote row i of P. For $P_i, P'_i \in \mathbb{R}^N$, we write $P_i < P'_i$ if $P_{ij} \leq P'_{ij}$ for all $j \in N$ and there is $k \in N$ such that $P_{ik} < P'_{ik}$.

Consider a financial network $F = (N, z, L, d)$. A *payment matrix* $P \in \mathscr{M}$ describes the mutual payments to be made by the agents, that is, P_{ij} is the

monetary amount to be paid by agent $i \in N$ to agent $j \in N$. Given a payment matrix $P \in \mathscr{M}$, the *asset value* $a_i(P)$ of agent $i \in N$ is given by

$$a_i(P) = z_i + \sum_{j \in N} P_{ji}.$$

Subtracting the payments made by an agent from the asset value yields an agent's equity. The *equity* $e_i(P)$ of an agent $i \in N$ is given by

$$e_i(P) = a_i(P) - \sum_{j \in N} P_{ij} = z_i + \sum_{j \in N} (P_{ji} - P_{ij}).$$

It follows immediately from the above expression that the sum over agents of their equities is the same as the sum over agents of their initial endowments. We have that

$$\sum_{i \in N} e_i(P) = \sum_{i \in N} z_i. \tag{1}$$

The analysis of financial networks is complicated because the mutual liability structure may result in contagion effects of default.

Our first aim is to define a clearing payment matrix. To do so, we define *feasible payments* of agent $i \in N$ as payments which belong to the image $d^i(\mathbb{R}_+)$ of the division rule d^i of agent i. A payment matrix is feasible if every row $i \in N$ of the matrix belongs to the feasible set of payments of agent i, that is, payments are made in accordance with the division rules. The set of *feasible payment matrices* $\mathscr{P}$ is therefore defined as

$$\mathscr{P} = \{P \in \mathscr{M} \mid \forall i \in N,\ P_i \in d^i(\mathbb{R}_+)\}.$$

The following definition of a clearing payment matrix is due to Csóka & Herings (2021). It extends the definition of Eisenberg & Noe (2001) for proportional division rules in a continuous setting. For a discrete setting with the smallest unit of account, see Csóka & Herings (2018).

Definition 2.1. The matrix $P \in \mathscr{M}$ is a *clearing payment matrix* of the financial network $F = (N, z, L, d)$ if it satisfies the following three properties:

1. *Feasibility:* $P \in \mathscr{P}$.
2. *Limited liability:* For every $i \in N$, $e_i(P) \geq 0$.
3. *Priority of creditors:* For every $i \in N$, if $P_i < L_i$, then $e_i(P) = 0$.

Limited liability requires all agents to end up with non-negative equity. Priority of creditors states that agents are only allowed to default if their equity is equal to zero.

Csóka & Herings (2021) prove the following result, which relates clearing payment matrices to the solution of a particular system of equations.

Theorem 2.2. *Let $F = (N, z, L, d)$ be a financial network. The payment matrix $P \in \mathscr{M}$ is a clearing payment matrix of F if and only if it solves the system of equations:*

$$P_{ij} = d_j^i(a_i(P)), \quad i, j \in N.$$

When calculating the clearing payment matrix as the solution to a system of equations, for every agent we take the value of the estate equal to the agent's asset value and next use the respective division rule to spend this asset value. Notice that agent $i \in N$ is treated as a claimant on its own estate $a_i(P)$ with a claim equal to $L_{ii} = 0$, so a clearing payment matrix P satisfies $P_{ii} = 0$.

3. MULTIPLICITY OF CLEARING PAYMENT MATRICES

We start by presenting two examples to show that clearing payment matrices need not be unique.

Example 3.1. We consider a financial network $F = (N, z, L, d)$ with three agents $N = \{1, 2, 3\}$, zero endowments $z = (0, 0, 0)$, and a liability matrix equal to

$$L = \begin{bmatrix} 0 & 4 & 8 \\ 8 & 0 & 4 \\ 4 & 8 & 0 \end{bmatrix}.$$

We examine the possible clearing payment matrices for the four common specifications of division rules: proportional, priority, constrained equal awards, and constrained equal losses.

We start with some general observations. Let P be a clearing payment matrix of F. By Definition 2.1, a clearing payment matrix satisfies limited liability, so, for every $i \in N$, it holds that $e_i(P) \geq 0$. Since by Equation (1) the sum over all agents of their equities is equal to the sum over all agents of their initial endowments, so $\sum_{i \in N} e_i(P) = \sum_{i \in N} z_i = 0$, it follows that, for every $i \in N$, $e_i(P) = 0$. The condition of priority of creditors in Definition 2.1 is therefore automatically satisfied and to find a clearing payment matrix, we

should therefore identify those payment matrices where all agents end up with zero equity while satisfying feasibility. A final observation is that in any clearing payment matrix the estates of the agents are all between 0 and 12.

Assume all agents use proportional division rules and let P be a clearing payment matrix of F. The estates of the agents satisfy the following system of equations:

$$E_i = \sum_{j \in N \setminus \{i\}} P_{ji} = \tfrac{2}{3} E_{i+1} + \tfrac{1}{3} E_{i-1}, \quad i \in N,$$

where we use the convention that agent 0 is identified with agent 3 and agent 4 with agent 1. It now follows from Gaussian elimination that $E_1 = E_2 = E_3$. Since estates of the agents are between 0 and 12, the set of clearing payment matrices when all agents use proportional division rules is given by

$$\mathscr{P}^{\text{prop}} = \{P \in \mathscr{M} \mid \exists E \in [0,12],\ \forall i \in N,\ P_i = \tfrac{1}{12} E L_i\}.$$

There is a one-dimensional, convex set of clearing payment matrices, ranging from no payments at all to full payments by all agents.

Next assume all agents use priority division rules, where the permutation is chosen such that larger liabilities have priority. The estates of the agents now satisfy the equations

$$E_i = \min\{E_{i+1}, 8\} + \max\{E_{i-1} - 8, 0\}, \quad i \in N. \tag{1}$$

Suppose not all estates are equal. Let $j \in N$ be such that $E_j < E_{j+1}$. It follows from the system of equations in (1) that $E_j \geq 8$, since the equation corresponding to E_j cannot hold with equality if $E_j < 8$ and $E_j < E_{j+1}$. We also have that

$$E_{j+1} = \min\{E_{j+2}, 8\} + \max\{E_j - 8, 0\} \leq 8 + E_j - 8 = E_j,$$

a contradiction to $E_j < E_{j+1}$. Consequently, it follows that all estates are equal, so $0 \leq E_1 = E_2 = E_3 \leq 12$.

The set of clearing payment matrices when all agents use priority division rules with the highest claim having priority is therefore given by

$$\begin{aligned} \mathscr{P}^{\text{prior}} \quad = \quad & \{P \in \mathscr{M} \mid \exists E \in [0,8],\ \forall i \in N,\ P_{i,i-1} = E \text{ and } P_{i,i+1} = 0\}, \\ & \cup \{P \in \mathscr{M} \mid \exists E \in [8,12],\ \forall i \in N,\ P_{i,i-1} = 8 \text{ and } P_{i,i+1} = E - 8\}. \end{aligned}$$

There is again a one-dimensional multiplicity of clearing payment matrices, ranging from no payments at all to full payments by all agents.

We now study the case of constrained equal award division rules. If the maximal estate across agents is less than or equal to 8, then the estates of the agents satisfy the following system of equations:

$$E_i = \sum_{j \in N \setminus \{i\}} P_{ji} = \tfrac{1}{2}E_{i+1} + \tfrac{1}{2}E_{i-1}, \quad i \in N.$$

It then follows that all estates must be equal, so $0 \leq E_1 = E_2 = E_3 \leq 8$. Any of these values of the estate generates a clearing payment matrix.

Next consider the case where the maximal estate across agents is strictly greater than 8. Let $j \in N$ be such that $E_j > 8$. Since $E_j = P_{j+1,j} + P_{j-1,j} \leq P_{j+1,j} + 4$, it holds that $P_{j+1,j} > 4$, so $E_{j+1} > 8$. We therefore find that all estates are strictly greater than 8. The system of equations becomes

$$E_i = E_{i+1} - 4 + 4 = E_{i+1}, \quad i \in N,$$

so solutions are given by $8 \leq E_1 = E_2 = E_3 \leq 12$. The set of clearing payment matrices when all agents use constrained equal awards division rules is therefore given by

$$\begin{aligned} \mathscr{P}^{\text{cea}} \;=\; & \{P \in \mathscr{M} \mid \exists E \in [0,8],\ \forall i \in N,\ P_{i,i-1} = \tfrac{1}{2}E \text{ and } P_{i,i+1} = \tfrac{1}{2}E\}, \\ & \cup \{P \in \mathscr{M} \mid \exists E \in [8,12],\ \forall i \in N,\ P_{i,i-1} = E - 4 \text{ and } P_{i,i+1} = 4\}. \end{aligned}$$

We again find a one-dimensional multiplicity of clearing payment matrices, ranging from no payments to full payments.

Finally, we examine the constrained equal losses division rules. If the maximal estate across agents is less than or equal to 4, then the estates of the agents satisfy the following system of equations:

$$E_i = \sum_{j \in N \setminus \{i\}} P_{ji} = E_{i+1}, \quad i \in N,$$

so solutions are given by $0 \leq E_1 = E_2 = E_3 \leq 4$. Consider next the case where at least one estate, say E_j, exceeds 4. Then agent j makes a payment greater than 4 to agent $j-1$, so E_{j-1} exceeds 4. It now follows that all estates exceed 4. We obtain the following system of equations:

$$E_i = 4 + \tfrac{1}{2}(E_{i+1} - 4) + \tfrac{1}{2}(E_{i-1} - 4) = \tfrac{1}{2}E_{i+1} + \tfrac{1}{2}E_{i-1},$$

so solutions are given by $4 \leq E_1 = E_2 = E_3 \leq 12$. The set of clearing payment matrices when all agents use constrained equal losses division rules is therefore given by

$$\begin{aligned}\mathscr{P}^{\text{cel}} &= \{P \in \mathscr{M} \mid \exists E \in [0,4],\ \forall i \in N,\ P_{i,i-1} = E \text{ and } P_{i,i+1} = 0\},\\ &\quad \cup \{P \in \mathscr{M} \mid \exists E \in [4,12],\ \forall i \in N,\ P_{i,i-1} = \tfrac{1}{2}E + 2 \text{ and}\\ &\quad P_{i,i+1} = \tfrac{1}{2}E - 2\}.\end{aligned}$$

Again a one-dimensional multiplicity of payment matrices results, which ranges from a least to a greatest clearing payment matrix. △

The next example shows that multiplicity of clearing payment matrices may depend on the division rules that are being used. This example also demonstrates the possibility of multiple clearing payment matrices when all agents have strictly positive endowments.

Example 3.2. We consider a financial network $F = (N, z, L, d)$ with three agents $N = \{1,2,3\}$, endowments $z = (3,6,7)$, and a liability matrix equal to

$$L = \begin{bmatrix} 0 & 6 & 4 \\ 12 & 0 & 5 \\ 0 & 0 & 0 \end{bmatrix}.$$

The highest possible asset value of agent 2 results when agent 1 pays the full liability $L_{12} = 6$ to agent 2, which leads to asset value $a_2(P) = z_2 + L_{12} = 6 + 6 = 12$ of agent 2. Since agent 2 has liabilities of 12 towards agent 1 and liabilities of 5 towards agent 3, agent 2 will always default and end up with zero equity due to priority of creditors, irrespective of the division rules in place.

We next examine the set of clearing payment matrices for the four most common specifications of division rules and start with the case of proportional division rules. From the system of equations presented in Theorem 2.2, we have that

$$\begin{aligned}P_{12} &= \min\{\tfrac{3}{5}(3 + P_{21}), 6\},\\ P_{21} &= \tfrac{12}{17}(6 + P_{12}).\end{aligned}$$

This system of equations has $P_{12} = 6$ and $P_{21} = 144/17$ as its unique solution. We find that the unique clearing payment matrix in the presence of proportional division rules and the resulting vector of equities are given by

$$P^{\text{prop}} \approx \begin{bmatrix} 0 & 6 & 4 \\ 8.47 & 0 & 3.53 \\ 0 & 0 & 0 \end{bmatrix} \qquad e(P^{\text{prop}}) \approx \begin{bmatrix} 1.47 \\ 0.00 \\ 14.53 \end{bmatrix}.$$

In the case of priority division rules with higher claims having priority, Theorem 2.2 leads to the following two equations:

$$\begin{aligned} P_{12} &= \min\{3+P_{21},6\}, \\ P_{21} &= 6+P_{12}. \end{aligned}$$

The unique solution is given by $P_{12}=6$ and $P_{21}=12$. We find that with priority division rules the unique clearing payment matrix and resulting vector of equities are given by

$$P^{\text{prior}} = \begin{bmatrix} 0 & 6 & 4 \\ 12 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad e(P^{\text{prior}}) = \begin{bmatrix} 5 \\ 0 \\ 11 \end{bmatrix}.$$

We continue with the examination of constrained equal awards division rules.

Suppose there is a clearing payment matrix such that the asset value of agent 1 is below 8. This implies $P_{21} = a_1(P) - z_1 < 8-3 = 5$. When using the constrained equal awards division rule, agent 2 only makes a payment to agent 1 less than 5 if the asset value of agent 2 is below 10. Theorem 2.2 yields the following equations:

$$\begin{aligned} P_{12} &= \tfrac{1}{2}(3+P_{21}), \\ P_{21} &= \tfrac{1}{2}(6+P_{12}). \end{aligned}$$

The only solution to this system of equations has $P_{12}=4$ and $P_{21}=5$, which is incompatible with an asset value of agent 1 below 8. Consequently, any clearing payment matrix results in an asset value of agent 1 of at least 8.

We next examine the existence of a clearing payment matrix where the asset value of agent 1 is at least equal to 8. To obtain such an asset value, the payment of agent 2 to agent 1 must at least be equal to 5. Under the constrained equal awards division rule, the asset value of agent 2 must then at least be equal to 10. The result of Theorem 2.2 gives rise to the following two equations:

$$\begin{aligned} P_{12} &= \min\{3+P_{21}-4,6\}, \\ P_{21} &= 6+P_{12}-5 = P_{12}+1. \end{aligned}$$

We find a continuum of solutions, with the value of P_{12} ranging between 4 and 6 and $P_{21}=P_{12}+1$. For every $E\in[8,10]$, we obtain a clearing payment matrix and resulting equities

$$P^{\text{cea}} = \begin{bmatrix} 0 & E-4 & 4 \\ E-3 & 0 & 5 \\ 0 & 0 & 0 \end{bmatrix} \qquad e(P^{\text{cea}}) = \begin{bmatrix} 0 \\ 0 \\ 16 \end{bmatrix}.$$

We conclude with the case of constrained equal losses division rules. Agent 1 has endowments equal to 3 and makes at least a payment of 2 to agent 2, so will achieve equal losses on the payments to agents 2 and 3. The asset value of agent 2 is at least equal to 8, so also agent 2 will achieve equal losses on the payments to agents 1 and 3. Theorem 2.2 gives rise to the following two equations:

$$\begin{aligned} P_{12} &= \min\{2+\tfrac{1}{2}(3+P_{21}-2),6\} = \min\{2\tfrac{1}{2}+\tfrac{1}{2}P_{21},6\}, \\ P_{21} &= 7+\tfrac{1}{2}(6+P_{12}-7) = 6\tfrac{1}{2}+\tfrac{1}{2}P_{12}. \end{aligned}$$

The unique solution is given by $P_{12} = 6$ and $P_{21} = 19/2$. We obtain the following market clearing payment matrix and corresponding equity:

$$P^{\mathrm{cel}} = \begin{bmatrix} 0 & 6 & 4 \\ 9.5 & 0 & 2.5 \\ 0 & 0 & 0 \end{bmatrix} \qquad e(P^{\mathrm{prior}}) = \begin{bmatrix} 2.5 \\ 0.0 \\ 13.5 \end{bmatrix}.$$

△

In Example 3.2, different division rules imply significantly different structural properties as far as clearing payment matrices are concerned. Constrained equal award division rules lead to a one-dimensional multiplicity of clearing payment matrices, whereas the clearing payment matrix is uniquely determined under the other division rules. Agent 1 defaults in almost all clearing payment matrices for constrained equal awards division rules, but not when any of the other division rules are used. Agent 2 fully defaults with respect to agent 3 under the priority division rule, fully pays the liability to agent 3 under constrained equal awards rules, whereas the claim of agent 3 on agent 2 is partially paid for under the other division rules.

A complete lattice is a partially ordered non-empty set in which every non-empty subset has a supremum and an infimum. In both Example 3.1 and in Example 3.2, the set of clearing payments matrices is a complete lattice. This turns out to be a general result as has been shown in Csóka & Herings (2021).

Theorem 3.3. *Let $F = (N, z, L, d)$ be a financial network. The set of clearing payment matrices of F is a complete lattice. In particular, there exists a least clearing payment matrix P^- and a greatest clearing payment matrix P^+.*

Eisenberg & Noe (2001) have shown Theorem 3.3 for the case of proportional division rules. Csóka & Herings (2018) prove a similar result in a discrete set-up.

4. CENTRALIZED CLEARING AS A PROGRAMMING PROBLEM

Csóka & Herings (2018) show in a discrete set-up that decentralized clearing results in the least clearing payment matrix. Ketelaars & Borm (2021) consider the continuous set-up and show that decentralized clearing processes converge to the least clearing payment matrix under mild conditions. Consider a decentralized clearing process where all agents simultaneously make the largest payments that are compatible with their cash at hand. In Example 3.1 all agents start with zero endowments, there are no positive feasible payments, and the decentralized clearing process stops at the least clearing payment matrix with zero payments. For the case of constrained equal awards division rules in Example 3.2, the decentralized clearing process is illustrated in Table 1.

z	L			P^1			P^2			...
3	0	6	4	0	1.5	1.5	0	3	3	...
6	12	0	5	3	0	3	3.75	0	3.75	...
7	0	0	0	0	0	0	0	0	0	...

P^{10}			...	P^-		
0	3.996	3.996	...	0	4	4
4.995	0	4.995	...	0	4	4
0	0	0	...	0	0	0

Table 1: The sequence of payment matrices using constrained equal awards division rules in Example 3.2.

In P^1, both agent 1 and agent 2 make equal payments to their creditors. Then, the new asset value of agent 1 becomes 3 and the new asset value of agent 2 becomes 1.5. In the next iteration, agents again make additional payments $P^2 - P^1$ in accordance with the constrained equal awards division rules. The payment matrices along the sequence show total payments made so far. P^{10} is rounded to three decimals. The process will take an infinite number of steps to converge to the least clearing payment matrix P^-.

As Examples 3.1 and 3.2 demonstrate, the amount of default can be significantly higher in the least clearing payment matrix than in the greatest clearing payment matrix. This triggers the natural question of whether it is possible to find the greatest clearing payment matrix. Since decentralized clearing processes end up in the least clearing payment matrix, doing so requires a

centralized approach. We show in this section that the greatest clearing payment matrix can be found by solving a particular maximization problem. For the division rules considered in this paper, this maximization problem can be written as a linear programming problem or an integer linear programming problem.

Throughout this section, $\mathbb{1}$ denotes a vector of ones of appropriate dimension. Theorems 4.1, 4.3, and 4.5 correspond to unpublished parts of Csóka & Herings (2017).

Theorem 4.1. *Let $F = (N, z, L, d)$ be a financial network. The greatest clearing payment matrix of F is the unique solution to the following maximization problem:*

$$\begin{array}{l} \max_{P \in \mathscr{P}} \sum_{i \in N} \sum_{j \in N} P_{ij}, \\ \textit{subject to} \\ z + P^\top \mathbb{1} - P\mathbb{1} \geq 0. \end{array} \tag{1}$$

Proof. Let P' be a solution to (1) and let some $i \in N$ be given. We show that $P'_i = d^i(a_i(P'))$.

If $P'_i = L_i$, then we have that

$$a_i(P') = z_i + \sum_{j \in N} P'_{ji} \geq \sum_{j \in N} P'_{ij} = \sum_{j \in N} L_{ij},$$

where the inequality follows from (1). From the definition of a division rule, it follows that $d^i(a_i(P')) = L_i$, so it holds that $P'_i = d^i(a_i(P'))$.

Consider the case $P'_i < L_i$. We show that $e_i(P') = 0$. Suppose $e_i(P') > 0$. Since $P' \in \mathscr{P}$ there exists $E' \in \mathbb{R}_+$ such that $P'_i = d^i(E')$. Since d^i is continuous and $e_i(P') > 0$ there exists $\varepsilon > 0$ such that

$$z_i + \textstyle\sum_{j \in N} P'_{ji} - \sum_{j \in N} d^i_j(E' + \varepsilon) \geq 0.$$

The payment matrix $P'' \in \mathscr{P}$ defined by

$$\begin{array}{lcll} P''_i & = & d^i(E' + \varepsilon), & \\ P''_j & = & P'_j, & j \neq i, \end{array}$$

satisfies the constraints in (1) and leads to a strictly higher value of the objective function than P', a contradiction. Consequently, it holds that $e_i(P') = 0$.

Since $P' \in \mathscr{P}$ there exists $E' \in \mathbb{R}_+$ such that $P'_i = d^i(E')$ and from $P'_i < L_i$ and the definition of a division rule, we have $\sum_{j\in N} d^i_j(E') = E'$. Since $e_i(P') = 0$, we therefore see that

$$E' = \sum_{j\in N} d^i_j(E') = \sum_{j\in N} P'_{ij} = z_i + \sum_{j\in N} P'_{ji} = a_i(P').$$

It follows that $P'_i = d^i(E') = d^i(a_i(P'))$.

We use Theorem 2.2 to conclude that P' is a clearing payment matrix.

Let P be any clearing payment matrix. By feasibility, it holds that $P \in \mathscr{P}$. By limited liability, it holds that, for every $i \in N$,

$$e_i(P) = z_i + \sum_{j\in N} P_{ji} - \sum_{j\in N} P_{ij} \geq 0.$$

Any clearing payment matrix therefore satisfies the constraints in (1). We see that P' is the clearing payment matrix with the largest sum of the payments made, so we can use Theorem 3.3 to conclude that P' must be the greatest clearing payment matrix. □

The maximization over $P \in \mathscr{P}$ in the program (1) guarantees that payments are feasible. The constraint ensures that no agent ends up with negative equity. The property that an agent is not allowed to default when having positive equity follows from the fact that the solution maximizes the objective function. Otherwise, it would be possible to increase the value of the objective function by having the defaulting agent make additional payments. The maximization of the objective function also guarantees that the greatest clearing payment matrix is selected.

When the financial network has proportional division rules, the greatest clearing payment matrix can be found as the solution to a linear programming problem. The following result has been shown in Eisenberg & Noe (2001). It follows as a special case of Theorem 4.1 when the feasibility constraint $P \in \mathscr{P}$ is replaced by explicit constraints that ensure payments are made according to proportional division rules.

Theorem 4.2. *Let $F = (N, z, L, d)$ be a financial network with proportional division rules. The greatest clearing payment matrix of F is the unique solution*

to the following linear programming problem:

$$\begin{array}{ll} \max_{P\in\mathbb{R}_+^{N\times N},\lambda\in\mathbb{R}_+^N} \sum_{i\in N}\sum_{j\in N} P_{ij}, & \\ \textit{subject to} & \\ P_{ij} = \lambda_i L_{ij}, & i,j\in N, \\ \lambda_i \leq 1, & i\in N, \\ z + P^\top \mathbb{1} - P\mathbb{1} \geq 0. & \end{array} \tag{2}$$

The first and second constraint in the linear program (2) guarantee that payments are proportional to the liabilities and at most equal to those liabilities. These constraints replace the requirement $P \in \mathscr{P}$ of the maximization problem (1). Demange (2018) uses a similar program to create a threat index by calculating the marginal effects of endowment increases.

Also for constrained equal awards division rules, we can replace the requirement $P \in \mathscr{P}$ of the program in (1) by a set of simple constraints. We define, for every $i \in N$, $\overline{L}_i = \max_{j\in N} L_{ij}$. Using Theorem 4.1, the following result follows in a straightforward way.

Theorem 4.3. *Let $F = (N, z, L, d)$ be a financial network with constrained equal awards division rules. The greatest clearing payment matrix of F is the unique solution P^+ to the following maximization problem:*

$$\begin{array}{ll} \max_{P\in\mathbb{R}_+^{N\times N},\lambda\in\mathbb{R}_+^N} \sum_{i\in N}\sum_{j\in N} P_{ij}, & \\ \textit{subject to} & \\ P_{ij} = \min\{\lambda_i, L_{ij}\}, & i,j\in N, \\ \lambda_i \leq \overline{L}_i, & i\in N, \\ z + P^\top \mathbb{1} - P\mathbb{1} \geq 0. & \end{array} \tag{3}$$

The program in (3) maximizes the total payments as made by the agents subject to three conditions. The first condition expresses that agent i pays all of its claimants the amount λ_i, except when λ_i would exceed the value of the claim. This yields the feasibility condition of clearing payment matrices under the constrained equal awards rule. The second condition serves to pin down a unique value of λ_i in all circumstances. It is possible to omit this constraint in the optimization problem, although one loses the property that λ_i is uniquely determined as well as the interpretation of λ_i as the highest

payment made by agent i. The third condition requires that no agent ends up with negative equity.

It is well-known that the constraint in (3) involving a minimum operator can be avoided by introducing binary decision variables q_{ij} for every $i, j \in N$. If $q_{ij} = 0$, then the payment P_{ij} is equal to L_{ij} and if $q_{ij} = 1$, then P_{ij} is equal to $\lambda_i \leq L_{ij}$. This leads to the following result.

Theorem 4.4. *Let $F = (N, z, L, d)$ be a financial network with constrained equal awards division rules. The greatest clearing payment matrix of F is the unique solution P^+ to the following integer linear programming problem:*

$$
\begin{array}{ll}
\max_{P \in \mathbb{R}_+^{N \times N}, \lambda \in \mathbb{R}_+^N, q \in \{0,1\}^{N \times N}} \sum_{i \in N} \sum_{j \in N} P_{ij}, & \\
\textit{subject to} & \\
P_{ij} \leq \lambda_i, & i, j \in N, \\
P_{ij} \leq L_{ij}, & i, j \in N, \\
P_{ij} \geq \lambda_i - \overline{L}_i(1 - q_{ij}), & i, j \in N, \\
P_{ij} \geq L_{ij} - \overline{L}_i q_{ij}, & i, j \in N, \\
\lambda_i \leq \overline{L}_i, & i \in N, \\
z + P^\top \mathbb{1} - P\mathbb{1} \geq 0. &
\end{array}
\tag{4}
$$

Proof. We show first that any $(P, \lambda, q) \in \mathbb{R}_+^{N \times N} \times \mathbb{R}_+^N \times \{0,1\}^{N \times N}$ satisfying the constraints in (4) is such that, for every $i, j \in N$, $P_{ij} = \min\{\lambda_i, L_{ij}\}$. We show next that for any $(P, \lambda) \in \mathbb{R}_+^{N \times N} \times \mathbb{R}_+^N$ satisfying the constraints in (3) there is $q \in \{0,1\}^{N \times N}$ such that (P, λ, q) satisfies the constraints in (4).

Let $(P, \lambda, q) \in \mathbb{R}_+^{N \times N} \times \mathbb{R}_+^N \times \{0,1\}^{N \times N}$ satisfy the constraints in (4). Let $i, j \in N$ be such that $q_{ij} = 0$. The constraints $P_{ij} \leq \lambda_i$, $P_{ij} \leq L_{ij}$, and $P_{ij} \geq L_{ij} - \overline{L}_i q_{ij} = L_{ij}$ imply $P_{ij} = \min\{\lambda_i, L_{ij}\}$. Let $i, j \in N$ be such that $q_{ij} = 1$. The constraints $P_{ij} \leq \lambda_i$, $P_{ij} \leq L_{ij}$, and $P_{ij} \geq \lambda_i - \overline{L}_i(1 - q_{ij}) = \lambda_i$ imply $P_{ij} = \min\{\lambda_i, L_{ij}\}$.

Let $(P, \lambda) \in \mathbb{R}_+^{N \times N} \times \mathbb{R}_+^N$ satisfy the constraints in (3). For every $i, j \in N$, if $P_{ij} < L_{ij}$, then define $q_{ij} = 1$, and if $P_{ij} = L_{ij}$, then define $q_{ij} = 0$. We show that (P, λ, q) satisfies the constraints in (4). Since $P_{ij} = \min\{\lambda_i, L_{ij}\}$, it follows that $P_{ij} \leq \lambda_i$ and $P_{ij} \leq L_{ij}$. If $P_{ij} < L_{ij}$, then $q_{ij} = 1$ and $P_{ij} = \lambda_i = \lambda_i - \overline{L}_i(1 - q_{ij})$. Clearly, it holds that $P_{ij} \geq 0 \geq L_{ij} - \overline{L}_i$. If $P_{ij} = L_{ij}$, then $q_{ij} = 0$ and $P_{ij} \geq 0 \geq \lambda_i - \overline{L}_i = \lambda_i - \overline{L}_i(1 - q_{ij})$. It also holds that $P_{ij} = L_{ij} = L_{ij} - \overline{L}_i q_{ij}$. □

To obtain desirable ormulations, we have treated all payments P_{ij} for $i, j \in N$ in the same way in the optimization problems (3) and (4). Of course, there

is no need to introduce explicit variables P_{ij} and q_{ij} when $L_{ij} = 0$ since we can simply substitute $P_{ij} = 0$.

Also for the constrained equal losses rule, we can replace the requirement $P \in \mathscr{P}$ of the program in (1) by a set of simple constraints. Using Theorem 4.1, we obtain the following result in a straightforward way.

Theorem 4.5. *Let $F = (N, z, L, d)$ be a financial network with constrained equal losses division rules. The greatest clearing payment matrix of F is the unique solution to the following maximization problem:*

$$\begin{array}{ll} \max_{P \in \mathbb{R}_+^{N \times N}, \mu \in \mathbb{R}_+^N} \sum_{i \in N} \sum_{j \in N} P_{ij}, & \\ \textit{subject to} & \\ P_{ij} = \max\{L_{ij} - \mu_i, 0\}, & i, j \in N, \\ \mu_i \leq \overline{L}_i, & i \in N, \\ z + P^\top \mathbb{1} - P\mathbb{1} \geq 0. & \end{array} \tag{5}$$

The program in (5) maximizes the total payments as made by the agents subject to three conditions. The first condition expresses that agent i pays all creditors the amount of their claim minus μ_i, except when μ_i exceeds the value of the claim. This corresponds to the feasibility condition of clearing payment matrices under the constrained equal losses rule. Similar to the case of constrained equal awards division rules, the second condition serves to pin down the value of μ_i. The only case where μ_i would not be uniquely determined without this constraint is when agent i does not make any payments in the greatest clearing payment matrix, a case that can only occur if $z_i = 0$ and i does not receive any payments from any of the other agents or if i does not have any creditors, both rather contrived situations. The third condition requires that no agent ends up with negative equity.

Similar to the case for constrained equal awards division rules, it is possible to avoid the constraint in (5) involving the maximum operator by introducing binary decision variables q_{ij} for every $i, j \in N$. If $q_{ij} = 0$, then the payment P_{ij} is equal to 0 and if $q_{ij} = 1$, then P_{ij} is equal to $L_{ij} - \mu_i$. This leads to the following result.

Theorem 4.6. *Let $F = (N, z, L, d)$ be a financial network with constrained equal losses division rules. The greatest clearing payment matrix of F is the*

unique solution P^+ to the following integer linear programming problem:

$$\begin{array}{ll}
\max_{P\in\mathbb{R}_+^{N\times N},\mu\in\mathbb{R}_+^N,q\in\{0,1\}^{N\times N}} \sum_{i\in N}\sum_{j\in N} P_{ij}, & \\
\textit{subject to} & \\
P_{ij} \geq L_{ij}-\mu_i, & i,j\in N,\\
P_{ij} \leq L_{ij}-\mu_i+\overline{L}_i(1-q_{ij}), & i,j\in N,\\
P_{ij} \leq \overline{L}_i q_{ij}, & i,j\in N,\\
\mu_i \leq \overline{L}_i, & i\in N,\\
z+P^\top \mathbb{1}-P\mathbb{1}\geq 0. &
\end{array} \tag{6}$$

Proof. We show first that any $(P,\mu,q)\in\mathbb{R}_+^{N\times N}\times\mathbb{R}_+^N\times\{0,1\}^{N\times N}$ satisfying the constraints in (6) is such that, for every $i,j\in N$, $P_{ij}=\max\{L_{ij}-\mu_i,0\}$. We show next that for any $(P,\mu)\in\mathbb{R}_+^{N\times N}\times\mathbb{R}_+^N$ satisfying the constraints in (5) there is $q\in\{0,1\}^{N\times N}$ such that (P,μ,q) satisfies the constraints in (6).

Let $(P,\mu,q)\in\mathbb{R}_+^{N\times N}\times\mathbb{R}_+^N\times\{0,1\}^{N\times N}$ satisfy the constraints in (6). Let $i,j\in N$ be such that $q_{ij}=0$. The constraints $P_{ij}\geq 0$, $P_{ij}\geq L_{ij}-\mu_i$, and $P_{ij}\leq \overline{L}_i q_{ij}=0$ imply $P_{ij}=\max\{L_{ij}-\mu_i,0\}$. Let $i,j\in N$ be such that $q_{ij}=1$. The constraints $P_{ij}\geq 0$, $P_{ij}\geq L_{ij}-\mu_i$, and $P_{ij}\leq L_{ij}-\mu_i+\overline{L}_i(1-q_{ij})=L_{ij}-\mu_i$ imply $P_{ij}=\max\{L_{ij}-\mu_i,0\}$.

Let $(P,\mu)\in\mathbb{R}_+^{N\times N}\times\mathbb{R}_+^N$ satisfy the constraints in (5). For every $i,j\in N$, if $P_{ij}>0$, then define $q_{ij}=1$, and if $P_{ij}=0$, then define $q_{ij}=0$. We show that (P,μ,q) satisfies the constraints in (6). Since $P_{ij}=\max\{L_{ij}-\mu_i,0\}$, it follows that $P_{ij}\geq L_{ij}-\mu_i$. If $P_{ij}>0$, then $q_{ij}=1$ and $P_{ij}=L_{ij}-\mu_i=L_{ij}-\mu_i+\overline{L}_i(1-q_{ij})$. Clearly, it holds that $P_{ij}=L_{ij}-\mu_i\leq L_{ij}\leq\overline{L}_i=\overline{L}_i q_{ij}$. If $P_{ij}=0$, then $q_{ij}=0$ and $P_{ij}=0\leq\overline{L}_i-\mu_i\leq L_{ij}-\mu_i+\overline{L}_i=L_{ij}-\mu_i+\overline{L}_i(1-q_{ij})$. It also holds that $P_{ij}=0=\overline{L}_i q_{ij}$. □

Finally, we turn to priority division rules. The following result follows immediately from Theorem 4.1.

Theorem 4.7. *Let $F=(N,z,L,d)$ be a financial network with priority division rules. The greatest clearing payment matrix of F is the unique solution P^+ to*

the following maximization problem:

$$\begin{aligned}
&\max_{P\in\mathbb{R}_+^{N\times N},E\in\mathbb{R}_+^N} \textstyle\sum_{i\in N}\sum_{j\in N} P_{ij},\\
&\textit{subject to}\\
&P_{ij} = \max\{0,\min\{E_i - \textstyle\sum_{k\in N|\pi(k)<\pi(j)} L_{ik}, L_{ij}\}\}, \quad i,j\in N,\\
&E_i \le \textstyle\sum_{k\in N} L_{ik}, \quad i\in N,\\
&z + P^\top \mathbb{1} - P\mathbb{1} \ge 0.
\end{aligned} \tag{7}$$

The program in (7) maximizes the total payments made by the agents subject to three conditions. The first condition expresses that agent i pays all creditors at most their claim or what is left after creditors having priority are paid off. The maximum makes sure that in case the latter amount is negative, no payments are made. This corresponds to the feasibility condition of clearing payment matrices under the priority rule. The second condition serves to pin down the value of E_i for solvent agents. The third condition requires that no agent ends up with negative equity.

The next result demonstrates that the problem of finding the greatest clearing payment matrix can be written as an integer linear programming problem as well in the case of priority division rules.

Theorem 4.8. *Let $F = (N,z,L,d)$ be a financial network with priority division rules. The greatest clearing payment matrix of F is the unique solution P^+ to the following integer linear programming problem:*

$$\max_{P\in\mathbb{R}_+^{N\times N},E\in\mathbb{R}_+^N,q\in\{0,1\}^{N\times N},r\in\{0,1\}^{N\times N}} \sum_{i\in N}\sum_{j\in N} P_{ij}, \tag{8}$$

subject to

$$P_{ij} \le L_{ij}, \quad i,j\in N, \tag{9}$$

$$P_{ij} \le E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik} + \sum_{k\in N} L_{ik}(1-q_{ij}), \quad i,j\in N, \tag{10}$$

$$P_{ij} \le \overline{L}_i q_{ij}, \quad i,j\in N, \tag{11}$$

$$P_{ij} \ge E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik} - \sum_{k\in N} L_{ik}(1-r_{ij}), \quad i,j\in N, \tag{12}$$

$$P_{ij} \ge L_{ij} - \overline{L}_i r_{ij}, \quad i,j\in N, \tag{13}$$

$$E_i \le \sum_{k\in N} L_{ik}, \quad i\in N, \tag{14}$$

$$z + P^\top \mathbb{1} - P\mathbb{1} \ge 0. \tag{15}$$

Proof. We show first that any $(P,E,q,r) \in \mathbb{R}_+^{N\times N} \times \mathbb{R}_+^N \times \{0,1\}^{N\times N} \times \{0,1\}^{N\times N}$ satisfying the constraints in (4.8) is such that

$$P_{ij} = \max\{0, \min\{E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik}, L_{ij}\}\}, \quad i,j \in N.$$

We show next that for any $(P,E) \in \mathbb{R}_+^{N\times N} \times \mathbb{R}_+^N$ satisfying the constraints in (7) there is $(q,r) \in \{0,1\}^{N\times N} \times \{0,1\}^{N\times N}$ such that (P,E,q,r) satisfies the constraints in (4.8).

Let $(P,E,q,r) \in \mathbb{R}_+^{N\times N} \times \mathbb{R}_+^N \times \{0,1\}^{N\times N} \times \{0,1\}^{N\times N}$ satisfy the constraints in (8). Let us fix some $(i,j) \in N \times N$. We distinguish three cases.

Case 1. $P_{ij} = 0$.

If $L_{ij} = 0$, then it clearly holds that

$P_{ij} = \max\{0, \min\{E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik}, L_{ij}\}\}$. Assume $L_{ij} > 0$. It follows from (13) that $r_{ij} = 1$. From (12) we obtain that $P_{ij} \geq E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik}$, so $P_{ij} = \max\{0, \min\{E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik}, L_{ij}\}\}$ as desired.

Case 2. $0 < P_{ij} < L_{ij}$.

It follows from (11) that $q_{ij} = 1$ and from (13) that $r_{ij} = 1$. We use (10) and (12) to conclude that $P_{ij} = E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik}$. We conclude that $P_{ij} = \max\{0, \min\{E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik}, L_{ij}\}\}$.

Case 3. $0 < P_{ij} = L_{ij}$.

It follows from (11) that $q_{ij} = 1$, so from (10) that $P_{ij} \leq E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik}$, and we conclude that $P_{ij} = \max\{0, \min\{E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik}, L_{ij}\}\}$.

Let $(P,E) \in \mathbb{R}_+^{N\times N} \times \mathbb{R}_+^N$ satisfy the constraints in (7). For every $i,j \in N$, we define $q_{ij} \in \{0,1\}$ and $r_{ij} \in \{0,1\}$ as follows. If $P_{ij} = L_{ij} = 0$, then define $q_{ij} = r_{ij} = 0$. If $P_{ij} = 0 < L_{ij}$, then define $q_{ij} = 0$ and $r_{ij} = 1$. If $0 < P_{ij} < L_{ij}$, then define $q_{ij} = r_{ij} = 1$. Finally, if $0 < P_{ij} = L_{ij}$, then define $q_{ij} = 1$ and $r_{ij} = 0$. We verify next that the inequalities in (9)–(13) are satisfied. To do so, we fix $(i,j) \in N \times N$ and distinguish four cases.

Case 1. $P_{ij} = L_{ij} = 0$.

The inequalities in (9)–(13) reduce to

$$\begin{aligned}
&0 \leq 0,\\
&0 \leq E_i + \textstyle\sum_{k\in N|\pi(k)\geq\pi(j)} L_{ik},\\
&0 \leq 0,\\
&0 \geq E_i - \textstyle\sum_{k\in N|\pi(k)<\pi(j)} L_{ik} - \sum_{k\in N} L_{ik},\\
&0 \geq 0.
\end{aligned}$$

These equalities are clearly satisfied, where the fourth inequality uses the fact that $E_i \leq \sum_{k\in N} L_{ik}$.

Case 2. $P_{ij} = 0 < L_{ij}$.

The inequalities in (9)–(13) reduce to

$$\begin{aligned} &0 \leq L_{ij},\\ &0 \leq E_i + \textstyle\sum_{k\in N|\pi(k)\geq\pi(j)} L_{ik},\\ &0 \leq 0,\\ &0 \geq E_i - \textstyle\sum_{k\in N|\pi(k)<\pi(j)} L_{ik},\\ &0 \geq L_{ij} - \overline{L}_i. \end{aligned}$$

Since $0 = P_{ij} = \max\{0, \min\{E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik}, L_{ij}\}\}$ and $L_{ij} > 0$, it follows that $E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik} \leq 0$, so the fourth inequality above holds. The other inequalities hold trivially.

Case 3. $0 < P_{ij} < L_{ij}$.

The inequalities in (9)–(13) reduce to

$$\begin{aligned} &P_{ij} \leq L_{ij},\\ &P_{ij} \leq E_i - \textstyle\sum_{k\in N|\pi(k)<\pi(j)} L_{ik},\\ &P_{ij} \leq \overline{L}_i,\\ &P_{ij} \geq E_i - \textstyle\sum_{k\in N|\pi(k)<\pi(j)} L_{ik},\\ &P_{ij} \geq L_{ij} - \overline{L}_i. \end{aligned}$$

Since $P_{ij} = \max\{0, \min\{E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik}, L_{ij}\}\}$ and $0 < P_{ij} < L_{ij}$, it follows that $P_{ij} = E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik}$. It is now easily verified that all inequalities above hold.

Case 4. $0 < P_{ij} = L_{ij}$.

The inequalities in (9)–(13) reduce to

$$\begin{aligned} &L_{ij} \leq L_{ij},\\ &L_{ij} \leq E_i - \textstyle\sum_{k\in N|\pi(k)<\pi(j)} L_{ik},\\ &L_{ij} \leq \overline{L}_i,\\ &L_{ij} \geq E_i - \textstyle\sum_{k\in N|\pi(k)<\pi(j)} L_{ik} - \sum_{k\in N} L_{ik},\\ &L_{ij} \geq L_{ij}. \end{aligned}$$

From $0 < L_{ij} = P_{ij} = \max\{0, \min\{E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik}, L_{ij}\}\}$ it follows that $E_i - \sum_{k\in N|\pi(k)<\pi(j)} L_{ik} \geq L_{ij}$. We have shown the second inequality above. The fourth inequality above follows from $E_i \leq \sum_{k\in N} L_{ik}$. The other inequalities are trivially true. □

We assumed in Theorems 4.3–4.8 that all agents use the same division rule. A closer inspection of the proofs reveals that this feature is not used. We can therefore obtain similar results if agents use heterogeneous division rules within the classes of proportional, constrained equal awards, constrained equal losses, and priority division rules.

5. CONCLUSION

We consider financial networks with perfectly liquid non-negative endowments, liabilities, and agent-specific bankruptcy rules. The set of clearing payment matrices is a complete lattice, so has a least and a greatest elements. We illustrate by means of examples that there can be an infinite number of clearing payment matrices and that the multiplicity of clearing payment matrices depends on the division rules that are in place. Previous research has shown that decentralized clearing leads to the selection of the least clearing payment matrix. We show how a centralized approach can be used to select the greatest clearing payment matrix. We present a programming approach to calculate the greatest clearing payment matrix. We also show that for proportional division rules, this programming problem can be written as a linear programming problem. For common division rules like constrained equal awards, constrained equal losses, and priority division rules, we show how the programming problem can be written as an integer linear programming problem.

There are many possibilities for further research. The Eisenberg & Noe (2001) model has been extended in various ways. For setups with default costs, see Rogers & Veraart (2013), Roukny et al. (2018), and Jackson & Pernoud (2020). Schuldenzucker et al. (2020) introduce credit default swaps and show how these can lead to multiplicity of clearing payment matrices. Cifuentes et al. (2005) analyze a related direct externality in financial networks, when agents' endowments also contain one illiquid asset. Defaulting agents sell the illiquid asset in a firesale, which reduces the other agents' value of endowments as well. In this setting, Amini et al. (2016) give conditions for uniqueness of the clearing payment matrix and the corresponding asset prices. Feinstein (2017) generalizes those conditions for multiple illiquid assets. If these conditions are not satisfied, there is again scope for multiplicity of clearing payment matrices. An examination of the trade-off between decentralized and centralized clearing is therefore highly relevant for the various extensions

of the baseline model.

References

Amini, H., Filipović, D., & Minca, A. (2016). Uniqueness of equilibrium in a payment system with liquidation costs. *Operations Research Letters*, *44*(1), 1–5.

Chatterjee, S., & Eyigungor, B. (2015). A seniority arrangement for sovereign debt. *American Economic Review*, *105*(12), 3740–65.

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. (2017). An axiomatization of the proportional rule in financial networks. *GSBE Research Memorandum 17/001*, *Graduate School of Business and Economics, Maastricht University, Maastricht, Netherlands*.

Csóka, P., & Herings, P. (2018). Decentralized clearing in financial networks. *Management Science*, *64*(10), 4681–4699.

Csóka, P., & Herings, P. (2021). Uniqueness of clearing payment matrices in financial networks. *GSBE Research Memorandum 21/014*, *Graduate School of Business and Economics, Maastricht University, Maastricht, Netherlands*.

Demange, G. (2018). Contagion in financial networks: A threat index. *Management Science*, *64*(2), 955–970.

Eisenberg, L., & Noe, T. H. (2001). Systemic risk in financial systems. *Management Science*, *47*(2), 236–249.

Feinstein, Z. (2017). Financial contagion and asset liquidation strategies. *Operations Research Letters*, *45*(2), 109–114.

Flores-Szwagrzak, K. (2015). Priority classes and weighted constrained equal awards rules for the claims problem. *Journal of Economic Theory*, *160*, 36–55.

Gallice, A. (2019). Bankruptcy problems with reference-dependent preferences. *International Journal of Game Theory*, *48*(1), 311–336.

Glasserman, P., & Young, H. P. (2016). Contagion in financial networks. *Journal of Economic Literature*, *54*(3), 779–831.

Groote Schaarsberg, M., Reijnierse, H., & Borm, P. (2018). On solving mutual liability problems. *Mathematical Methods of Operations Research*, *87*(3), 383–409.

Jackson, M. O., & Pernoud, A. (2020). Credit freezes, equilibrium multiplicity, and optimal bailouts in financial networks. *arXiv preprint arXiv:2012.12861*.

Jackson, M. O., & Pernoud, A. (2021). Systemic risk in financial networks: A survey. *Annual Review of Economics*, *13*, 171–202.

Ketelaars, M., & Borm, P. (2021). On the unification of centralized and decentralized clearing mechanisms in financial networks. *CentER Discussion Paper Series 2021-*

015, *Tilburg University, Tilburg*, 1–41.

Moulin, H. (2000). Priority rules and other asymmetric rationing methods. *Econometrica*, *68*(3), 643–684.

Rogers, L. C., & Veraart, L. A. (2013). Failure and rescue in an interbank network. *Management Science*, *59*(4), 882–898.

Roukny, T., Battiston, S., & Stiglitz, J. E. (2018). Interconnectedness as a source of uncertainty in systemic risk. *Journal of Financial Stability*, *35*, 93–106.

Schuldenzucker, S., Seuken, S., & Battiston, S. (2020). Default ambiguity: Credit default swaps create new systemic risks in financial networks. *Management Science*, *66*(5), 1981–1998.

Thomson, W. (2003). Axiomatic and game-theoretic analysis of bankruptcy and taxation problems: A survey. *Mathematical social sciences*, *45*(3), 249–297.

Thomson, W. (2013). Game-theoretic analysis of bankruptcy and taxation problems: Recent advances. *International Game Theory Review*, *15*(03), 1340018.

Thomson, W. (2015). Axiomatic and game-theoretic analysis of bankruptcy and taxation problems: An update. *Mathematical Social Sciences*, *74*, 41–59.

Journal of Mechanism and Institution Design
ISSN: 2399-844X (Print), 2399-8458 (Online)
DOI:10.22574/jmid.2022.12.003

CENTRALIZED REFUGEE MATCHING MECHANISMS WITH HIERARCHICAL PRIORITY CLASSES

Dilek Sayedahmed

WISIR, University of Waterloo, Canada
dilek.sayedahmed@uwaterloo.ca

ABSTRACT

This study examines the refugee reallocation problem by modeling it as a two-sided matching problem between countries and refugees. Based on forced hierarchical priority classes, I study two interesting refugee matching algorithms to match refugees with countries. Axioms for fairness measures in resource allocation are presented by considering the stability and fairness properties of the matching algorithms. Two profiles are explicitly modeled—country preferences and forced prioritization of refugee families by host countries. This approach shows that the difference between the profiles creates blocking pairs of countries and refugee families owing to the forced hierarchical priority classes. Since the forced priorities for countries can cause certain refugees to linger in a lower priority class in every country, this study highlights the importance of considering refugees' preferences. It also suggests that a hierarchical priority class-based approach without category-specific quotas can increase countries' willingness to solve the refugee reallocation problem.

Keywords: Stability, deferred acceptance algorithm, refugee studies.

JEL Classification Numbers: D47, C78, D78, I30.

I am deeply grateful to my Ph.D. thesis advisor Szilvia Pápai. I would like to thank Will Jones and Alexander Teytelboym for their great inspiration and guidance. I am grateful to Péter Biró for his helpful feedback as my doctoral thesis examiner in 2021, as well as Samson Alva for his comments at the 2021 Virtual Conference on Social Choice Theory and Applications. I sincerely thank the three anonymous referees and the editor of the Journal, as well as the audience at the 2019 University of Sherbrooke GREDI/CREATE/CIREQ Workshop for their truly helpful comments. Financial support from Fonds de recherche du Québec Société et culture (FRQSC) is gratefully acknowledged. Any errors are my own.

1. INTRODUCTION

In the international refugee regime, it has posed a real challenge to finding a solution to the European refugee crisis. European countries have been reluctant to participate in the responsibility-sharing of resettling refugee families. Given this context, this study investigates the problem of resource allocation. Furthermore, this study draws inspiration from Abdulkadiroğlu & Sönmez (2003), in which they formulate the school choice problem as a mechanism design problem. They propose two competing mechanisms—the student optimal stable mechanism (Gale & Shapley, 1962) and the top trading cycles mechanism—each providing a solution to the school choice problem; see Shapley & Scarf (1974), Pápai (2000), and Abdulkadiroğlu & Sönmez (2003) on the top trading cycle mechanism. In this study, we reformulate the refugee reallocation problem as a mechanism design problem to address to the critical international refugee management issues. I define a two-sided matching problem with refugee and country preferences as a *country acceptance problem*. The one-sided version of this matching problem would be with refugees' preferences and countries' predetermined priority rankings. Consequently, based on the mechanism design literature, these problems are similar to college admissions and school choice problems, respectively.

Here two questions arise naturally. First, how can we model the interplay between forced priority classes and country preferences, that is, how can we capture the conflict between a one-sided and a two-sided model in the refugee reallocation context? Although the preferences of countries seem to be frowned upon in political debates, the global reality indicates that for a stable responsibility-sharing, preferences of countries tend to be crucial and influential. Second, what type of stability measures could be outlined for such a problem? I address the country acceptance problem by designing two matching algorithms based on forced hierarchical priority classes. These algorithms could be implemented as centralized refugee matching systems that match refugee families to countries. In this study, the term "refugee" is used in reference to a refugee family. In designing a centralized matching system that ensures the inseparability of refugee families that do not wish to be separated, this study assumes the implementation of a clearinghouse to accept preference submissions from households, namely, refugee families. All participating countries in the clearinghouse treat refugee households as a single refugee family unit.

Moreover, I conduct an axiomatic allocation analysis by focusing on the stability and fairness properties of the matching mechanisms. I explicitly model and analyze two profiles—countries' preferences, and the prioritization of refugee families imposed on host countries. Having two kinds of ranking profiles for countries, a *forced priority profile* and a *preference profile*, allows us to capture the difference between the two profiles. This, thereby, creates blocking pairs of countries and refugees owing to the forced hierarchical priority classes. The forced priority profile is a joint master list that is designed according to the United Nations High Commissioner for Refugees' (UNHCR) principles. It is applied to all countries, which may conflict with some countries' preferences. Similar models have been used in the resident allocation problem with distributional constraints in the computer science and artificial intelligence literature; see e.g., Goto et al. (2016).

In this context, I recognize the importance of investigating the weakening of the stability and fairness axioms. A stable mechanism may no longer satisfy the standard stability of the literature concerning country preferences, leading to potential blocking pairs. Therefore, I contribute to the literature by studying weaker stability and fairness axioms in order to determine the type of stability and fairness properties that hold in this setting. The motivation behind this is twofold. First, I account for the fact that countries have their preferences. Second, the UNHCR has strict humanitarian guidelines and principles for refugee settlement, which are a pivotal consideration when designing a centralized refugee matching mechanism (Assembly et al., 1951; Szobolits & Sunjic, 2007). Based on these guidelines, I impose priority classes on participating countries that force countries to change their preference rankings. This, however, leads to a deviation from the goal of a fair refugee allocation. Owing to these forced priorities for the countries giving certain refugees a priority in *each* country, I recognize the need for and the importance of considering refugee preferences. Since priority classes are forced in all countries, every refugee in the first priority class (PC) is always prioritized over others. Thus, in my proposed system, a refugee in a lower category remains in that category. Please note that the abbreviation "PC" is used for "priority class" throughout this study.

Considering the refugees who linger in the lower forced priority classes, I focus on two different forms of priority profiles to give these refugees an additional chance to improve their ranking. In my first form, namely the top prioritization mechanism, I provide these refugees with a higher probability

to be matched with their top-ranked country. In my second form, I present them with a better chance of being paired with their Deferred-Acceptance-matched country. It must be noted that, throughout this study, I refer to Gale & Shapley (1962)'s refugee-proposing Deferred-Acceptance mechanism as "the DA," which is applied to the refugee and country preferences. The second form is defined as the DA-match prioritization mechanism. Moreover, unlike the first form, I find that the DA-match is the best matching that the refugees in lower priority classes can acquire in a stable matching scenario. This shows the importance of prioritizing these refugees in their DA-matched countries.

For countries, the prioritization of these refugees in their DA-matched countries is more compelling, given that each country's priorities undergo fewer quota-based modifications than those in the first form of a priority profile based on refugees' top-ranked countries. Under the first form of a priority profile, it may not be necessary for some countries to modify the priority orders, despite it being mandatory for others to do so, and to do so to a greater extent. Let us consider Germany—the desired country for refugee settlement. When Germany becomes the top choice for a large number of refugee families, the country will move several refugees to its top priority class, ranking them according to its own point system. This system represents Germany's preferences. Consequently, a favored country such as Germany can make several changes to its priority ordering. This will cause the country to deviate significantly from the forced priority compared with a less popular country among refugees.

This study's contribution lies at the intersection of the matching theory and refugee studies through multiple channels. First, although countries may have clear preferences for refugees, they are not required to be familiar with all the predispositions over the entire set of refugees to run the mechanisms designed in this study. Since these mechanisms would require the countries to submit preferences over refugees in the same priority class, it would enable the countries to implement the mechanisms more efficiently. Second, challenges may arise from the imposition of type-specific (e.g., PC-specific), set-aside reserve quotas on countries in a refugee allocation setting. Hence it may not be well-accepted by the countries. However, a hierarchical priority class-based approach without category-specific set-aside reserve quotas may be more acceptable and induce more countries to willingly solve the refugee reallocation problem. This approach would persuade more countries to participate in a centralized refugee matching mechanism. This study's find-

ings have other important, decisive policy implications. For example, they can be applied to centralized college admissions, the design of public-school choice systems, and an immigration process characterized by a more effective priority-based than a category-specific reserve quota system.

Literature Review

Top prioritization mechanism may seem part of the first choice maximizing (FCM) mechanisms of Dur et al. (2018). The motivation of their study of FCM mechanisms is the common focus on first choices in school choice markets and the popularity of variants of the Boston mechanism (BM) in practice. BM is popular and it has an intuitive way of attempting to maximize first choices. However, BM was not part of the motivation of this study. The top prioritization mechanism was motivated by the need to identify the means of helping refugees who face the risk of being stuck in a low priority class in all countries. When searching for a way to help move these refugees up to a higher priority class, a natural starting point was to look at these refugees' most-preferred countries. BM can be understood as DA with a modified priority profile, where one first sorts agents according to their rank of the object (agents who rank the object first are in the top PC, agents who rank the object second are in the second PC, etc.) and, within each PC, the agents are ordered according to the original priorities of the object.

In contrast, the top prioritization mechanism is simply the DA applied to a newly adjusted forced priority profile in which refugees are moved up to top PC of their most-preferred countries. Moreover, when these refugees are promoted to top PC, they are still ranked according to country preferences within the top PC. Meanwhile, the first choices first (FCF)-algorithm of Dur et al. (2018) is a procedure in which at step one each student applies to her respective first choice school; and each school accepts applicants into open seats according to priorities until there are no more applicants or all seats are filled. In step two, rejected students are matched to open seats by an arbitrary procedure but without changing the matchings that were made in step one. Furthermore, the top stability axiom of this study, which is a weakened stability axiom that allows for blocking pairs of refugees and countries that are not their top choice, is Dur et al. (2018)'s first choice-stability. A matching is first choice-stable if no student forms a blocking pair with her first choice.

Credible stability is one of the other weak stability axioms examined in

this study. It allows for salient blocking pairs that are not matched under the DA in the refugee allocation setting. It is an interesting concept as it can potentially improve upon the DA refugee-optimal solution, although without being strategyproof. This axiom may remind the reader of the Optimal Priority DA proposed in Biró & Gudmundsson (2021). In their study, they examine the complexity of finding Pareto-efficient allocations of highest welfare in a school choice context. The Optimal Priority DA is a DA executed on the instance with priorities adjusted to the welfare-maximizing allocation based on minimized distance. The idea of the Optimal Priority DA is that the authors take an optimal matching (e.g. one that minimises the total travel distance for the students) and they provide top priority to all students at the schools where they are assigned in this socially optimal solution. Then they apply a standard DA taking into account the students' preferences and the further priorities at the schools. While Optimal Priority DA starts by computing the welfare-maximizing allocation, the DA-match prioritization mechanism depends on the initial DA allocations based on country and refugee preferences. Moreover, since the DA-match prioritization mechanism weakly improves on the DA and is manipulable, domains studied in Kesten & Kurino (2019) are of interest to the readers. They identify maximal domains on which strategy-proof mechanisms dominating DA exist. The motivation of their paper is different than this study, as their main goal is the improvement of the DA and examining the level of strategy-proofness loss as a result. Meanwhile, in this study, the DA-match prioritization mechanism is designed as an alternative channel to give refugees, who face the risk of being stuck in a lower PC, an improved rank by moving them up to a higher PC. The Pareto-improvement result that came with the DA-match prioritization mechanism's design was a pleasant surprise.

Since the two mechanisms of top and DA-match prioritization in this study are manipulable, they are of interest with respect to the mechanisms studied in Pathak & Sönmez (2013). Their approach to studying a mechanism's vulnerability to manipulation is to characterize domains under which the mechanism is not manipulable. They develop a rigorous methodology to compare mechanisms based on their vulnerability to manipulation. Moreover, the weak stability of Pathak & Sönmez (2013) seems similar to the top stability of this study. Their weak stability is a relaxation of stability; for example, students are allowed to block matchings only with their top choice schools. Meanwhile, under the top stability of this study, which is also a relaxation of stability, such

blocking pairs are the ones that are not allowed.

The remainder of this paper proceeds as follows. Section 2 provides a background on the refugee reallocation crisis, elaborates on this study's motivation, and outlines the centralized system proposed along with the UNHCR-mandated priority classes. Section 3 presents the model. Section 4 investigates how we can weaken stability in the context of hierarchical UNHCR-mandated priority classes. This section also provides the basic definitions of the axioms and related theorems. Sections 5 and 6 present the two weak stability axioms that consider the top-ranked and DA-matched countries when modifying priorities that improve the chance of refugees in the lower priority classes. Section 7 concludes this study.

2. BACKGROUND AND MOTIVATION

The number of global refugees is currently at its highest level since the Second World War, and Europe has attracted the predominant mass of refugees (Alfred, 2015). This scenario shows the European Union's (EU's) inconsistency and poorly coordinated response to refugees fleeing the collapse of Syria. This issue has attracted several studies seeking solutions to the refugee allocation problems at the international and local levels. Considering the local refugee match within a country, studies have presented several obstacles to successful refugee integration. Andersson & Ehlers (2020) focus on the problem of finding housing for refugees after their resettlement to an EU country. They propose an easily implementable algorithm for Sweden that finds a stable maximum matching. Further, Bansak et al. (2018) and Delacrétaz et al. (2019) focus on different aspects of the refugee allocation problem: the former on the optimization of refugee preferences and the latter on family size.

In the context of the international refugee allocation problem, there have been several calls for revising or replacing the Dublin Regulation, particularly its requirement that the first EU country that gives asylum undertake the responsibility of processing the asylum seekers' claims (Giuffre & Costello, 2015; Koser, 2011). The current decentralized system, which assigns this responsibility to the first-arrival country, has been unfair to border countries such as Greece, Italy, or Hungary. Subsequently, this system has also been responsible for creating chaos and tragedy. One article criticizes the European states for playing "pass-the-parcel with human lives (Jones & Teytelboym, 2017)." I agree with Jones & Teytelboym (2017) that "it has never been clearer

that a new deal on responsibility-sharing within Europe is needed to replace the Dublin Regulation."

It is crucial to implement a centralized matching system alongside or, ideally, instead of the current system of the Dublin Regulation. A new centralized system would allow a refugee family to apply for protection in more than just one country, thus not binding them to a single application to a particular country. Under the current system, refugees take the gamble of deciding where to apply, and countries cannot evaluate and choose from a large pool of applicants.

A centralized mechanism would allow all refugees to apply to a single system from any embassy, and this could benefit both refugees and countries (Jones & Teytelboym, 2017, 2016). Just as a centralized mechanism could help refugees be better off by helping them avoid dangerous cross-border journeys, a centralized mechanism could also enable countries to gain more control than they currently possess in deciding who settles within their borders. This can be ensured by allowing countries to assign their preference ranking on which refugees they wish to accept, similar to refugees citing their preferences for countries (Jones & Teytelboym, 2017, 2016).

Furthermore, a centralized matching system would allow refugees to apply for protection in several countries while allowing the countries to compete for refugees. This system would require the refugees to make only a single claim for asylum to a single centralized body while simultaneously specifying their country preference. When the countries approach the clearing house with a quota and ranking of refugees, they are willing to accept that the system can be implemented to match the refugees to the countries. After this matching process, it would be crucial to implement this match, that is, refugees are granted refugee status and permitted to settle in the country to which they have been matched. A centralized system would allow refugees to apply for asylum with *every* participating country, and they can, in principle, submit it remotely. This submission would include the regional processing centers in the Middle East and North Africa (Jones & Teytelboym, 2017, 2016). The core advantage of the system is that it provides refugee families with confidence that a fair and effective system will grant them protection, and they will be less likely to take the risk of a dangerous crossing.

Moreover, the UNHCR's *Convention and protocol relating to the status of refugees* (Assembly et al., 1951) is both a status- and rights-based instrument. It is underpinned by several fundamental principles, such as safety and protec-

tion, non-discrimination, non-penalization, and *non-refoulement*. Convention provisions, for example, are to be applied without discrimination as to race, religion, or country of origin. Developments in international human rights law also reinforce the principle that the Convention be applied without discrimination based on sex, age, disability, sexuality, or other prohibited grounds of discrimination (Assembly et al., 1951).

The UNHCR document states the following:[1]

"The conference, considering that the unity of the family—the natural and fundamental group unit of society—is an essential right of the refugee, and that such unity is constantly threatened, and noting with satisfaction that, according to the official commentary of the *ad hoc* committee on 'Statelessness and Related Problems,' the rights granted to a refugee are extended to members of his family, recommends governments to take the necessary measures for the security and protection of the refugee's family, especially with a view to:

1. Ensuring that the unity of the refugee's family is maintained particularly, in cases where the head of the family that has been waiting for admission has fulfilled the necessary conditions for admission to a particular country,

2. The protection of refugees who are minors, in particular unaccompanied children and girls, with special reference to guardianship and adoption (Assembly et al., 1951)."

The Forced Priority Class Hierarchy: A Proposal

A centralized system aims to allocate refugees within that system to a country where they are most likely to flourish during their time of residency, without causing further immigration spillover to other countries (Jones & Teytelboym, 2017, 2016). It is essential to ensure that such a system excludes discriminatory categories and focuses on the categories of vulnerability, the suitability for integration, and the presence of family. Hence, I propose *forced hierarchical priority classes* of refugees for countries based on the UNHCR's 1951

[1] For the full texts of the UNHCR 1951 documents, please see the Convention and Protocol relating to the status of refugees from the UN General Assembly in Geneva (Assembly et al., 1951).

convention and protocol for refugees (Assembly et al., 1951). Throughout this study, I refer to this UNHCR-mandated PC hierarchy of refugees as the *forced priority classes*. A forced PC of refugees is a subset of the entire set of refugees a country must consider as part of the PC hierarchy, which is the same for all participating countries. The forced PC hierarchy is imposed exogenously as part of the centralized system. The proposed forced priority profile for countries is then as follows:

- *Priority Class I:* Refugee families in war zones and with the longest waiting period
- *Priority Class II:* Refugee families in war zones
- *Priority Class III:* Refugee families with the longest waiting period
- *Priority Class IV:* Other refugee families

Given the initial theoretical approaches to complicated real-life problems, such as the refugee crisis, it is a requisite to start the analysis with a static model. Therefore, for its theoretical tractability, this study chose to take a static approach. The key motivations for this study are the Syrian refugee crisis and the reallocation problems resulting from the Syrian war. The Syrian crisis has led to an influx of refugees into other countries, which creates a refugee pool. Thus, we can conduct this exercise repeatedly. However, in the school choice, there are natural time periods. For example, the implementation of school choice would require considering each academic year. However, it is unclear whether the same process can be applied to refugee settlement with a sudden influx of refugees. Nonetheless, since this exercise can be implemented, for instance, every three months when there is a crisis, we can consider a dynamic approach. By heeding this approach, a static approach would also be crucial to address a refugee crisis that creates a sudden large pool of refugee families requiring immediate allocation.

Usually, dynamic models can better capture the essence of the problem under study than static models do. In dynamic models, agents nn arrive at different time periods and, sometimes, continuously; matching takes place at different time periods, and the matched agents leave the market, and new agents arrive. When this is the case, we lose some desirable properties in a static matching. Thus, we can turn the matching environment of refugee allocation into a static environment and retain desirable properties of stability and

efficiency. When possible, it would be ideal to array agents from both sides to conduct repeated static matchings. This would be an optimal approach in an environment with a crisis and is better than the dynamic approach of performing repeated matchings in each time period as refugees enter and leave the market. Hence, in the case of a sudden influx of refugee families, the static approach can be adapted to facilitate matching these participants, who are fixed at the moment of the crisis when the matching takes place.

This study contributes to the literature as the model explicitly allows for the potential real-world "discrepancy" between what countries actually want to do (i.e., country preferences) and what countries are forced to do by law (i.e., countries' forced priorities). By working with two profiles for countries, I capture the "compromise" between countries' *actual* preferential ranking of refugees in a world with no special considerations and their *forced* priority hierarchy based on the UNHCR's humanitarian laws and principles. Forced priorities are used for school choice by forcing the local authorities to use the main priority categories, such as living in the catchment area and/or having a sibling in the school. This study introduces the perspective of compromise between desired and forced preferences to the refugee reallocation problem by making it the central topic of the policy-making discussion. Moreover, this study explicitly lays out this compromise and its potential implications for fairness. In the refugee reallocation context, country preferences are negatively reviewed due to discrimination issues. Therefore, the hierarchical classes would aid in cutting down significantly on this kind of discrimination in the refugee setting by imposing the UNHCR-based priority classes, which makes country preferences much less important for the matching outcome. However, on the other hand, refugees face the risk of being stuck in lower priority classes in all countries. This study contributes by attempting to find a middle ground between these two forces by designing mechanisms that help those refugees who might face the risk of remaining in lower priority classes.

3. THE MODEL

Definition (Country Acceptance Problem). A problem consists of the following:

1. A finite set of refugees $\mathscr{R} = \{r_1, r_2, ..., r_{|\mathscr{R}|}\}$.

2. A finite set of countries $\mathscr{C} = \{c_1, c_2, ..., c_{|\mathscr{C}|}\}$.

3. A quota vector $q = (q_{c_1}, ..., q_{c_{|\mathscr{C}|}})$, where q_c is the capacity (the number of residency permits) of $c \in \mathscr{C}$.

4. Preference profile of refugees $P = (P_{r_1}, P_{r_2}, ..., P_{r_{|\mathscr{R}|}}) = (P_r)_{r\in\mathscr{R}}$, where P_r is the strict preference of refugee $r \in \mathscr{R}$ over $\mathscr{C}$.

5. Preference profile of countries $\succ = (\succ_{c_1}, \succ_{c_2}, ..., \succ_{c_{|\mathscr{C}|}}) = (\succ_c)_{c\in\mathscr{C}}$, where $\succ_c$ is the strict preference of country $c \in \mathscr{C}$ over $\mathscr{R}$, based on its country-specific point system.[2]

A finite set of refugees $\mathscr{R}$ is with a *fixed partition*, such that $\mathscr{R} = \mathscr{R}_1 \cup \mathscr{R}_2 \cup ... \cup \mathscr{R}_T$ and $\mathscr{R}_t \cap \mathscr{R}_{t'} = \emptyset$, for any $t, t' \in \{1, ..., T\}$). The T number of *forced hierarchical priority classes* of the partition are class $\mathscr{R}_1$, class $\mathscr{R}_2$, and so on until class $\mathscr{R}_T$. These are the forced priority classes introduced and discussed in the previous section.
As a primitive version of the model, I also define the following:

Definition (Forced Priorities). Countries' enforcing priority profile $\pi^E = (\pi^E_c)_{c\in\mathscr{C}}$ is based on the fixed partition $\mathscr{R}_1, ..., \mathscr{R}_T$ of the refugees into the forced priority classes.

Let $t, t' \in \{1, ..., T\}$ such that $t \leq t'$ and let $r_i \in \mathscr{R}_t$, $r_j \in \mathscr{R}_{t'}$. Then, for all $c \in \mathscr{C}$,

a. If $t = t'$, then $r_i \; \pi^E_c \; r_j$ if and only if $r_i \succ_c r_j$.

b. If $t < t'$, then $r_i \; \pi^E_c \; r_j$.

For item (a) above, observe that each country's rankings within each PC are country-specific preferences. For any $c \in \mathscr{C}$ with $r_i \succ_c r_j$, if r_i and r_j are prioritized within the *same* class, then *within* that PC, country c's ranking will also be $r_i \; \pi^E_c \; r_j$, following country c's preference ranking. For item (b) above, observe that if one refugee is in a higher PC than another refugee, then this implies that, according to π^E, the refugee in the higher priority class will be

[2] Unacceptable countries are not allowed in the model, in which case each refugee family $r \in \mathscr{R}$ has strict preferences over $\mathscr{C}$, where a country c is never ranked below r. Similarly, unacceptable refugees are not considered in the model. Hence, each country $c \in \mathscr{C}$ has strict preferences over $\mathscr{R}$, where a refugee r is never ranked below c.

strictly prioritized over the refugee who is in a lower PC.

Remark. In partitioned forced priorities, each member of the partition corresponds to an exogenously imposed PC. The priority classes are forced across all participating countries; each forced PC consists of the same set of refugees for all participating countries.

In summary, a many-to-one *country acceptance problem*, where each refugee family can be matched to a maximum of one host country and each country can admit a maximum of q_c refugee families, is defined as the tuple

$$\langle \mathscr{R} = \bigcup_{t=1}^{T} \mathscr{R}_t, \mathscr{C}, (q_c)_{c\in\mathscr{C}}, P, \succ \rangle.$$

When all other parameters, except that of the refugees' preference profile P, are fixed, I refer to P as a country acceptance problem. Let $\mathscr{P}$ denote the set of country acceptance problems.

To simplify the exposition, I assume that for all $c \in \mathscr{C}$, $|\mathscr{R}| > q_c$. This assumption is easily satisfied in any application and allows the rejection of the case where there are few refugees. I also perform the following notation: Let W_r denote the weak preferences of refugee $r \in \mathscr{R}$ associated with P_r.[3] Since preferences are strict, $c\ W_r\ c'$ means that either $c\ P_r\ c'$ or $c = c'$. The preferences of a coalition $L \subseteq \mathscr{R}$ in P are denoted by P_L. Finally, I denote the preference profile of all the refugees, except for r by P_{-r}, and the preference profile of all refugees, except the ones in coalition L by P_{-L}.

For simplicity, I assume that countries have *responsive preferences* over refugee families. This means that relatively speaking, refugee families are not complements in the countries' preferences. Thus, preferences over sets of refugee families can be interpreted as a natural extension of preferences over individual refugees.

Definition (Responsive Preferences). For any set of refugee families, $\mathscr{Z} \subset \mathscr{R}$ with $|\mathscr{Z}| \leq q_c$ and any refugee family r and r' in $\mathscr{R} \setminus \mathscr{Z}$,

- $\mathscr{Z} \cup \{r\} \succ_c \mathscr{Z} \cup \{r'\}$ if and only if $r \succ_c r'$.

[3] Please note that symbol W is used to denote weak preference relation for refugees. This is done to avoid confusion between the common notation that is used for weak preference (R) associated with P and set of refugees $\mathscr{R}$.

It must be noted that the notation has been slightly abused in the above definition. To indicate preferences over sets, $\succ_c$ is used; it is normally used for showing preferences over singletons in $\mathscr{R}$. Responsive preferences are a natural extension of preferences over individuals to preferences over sets. This property does not give a complete ordering of all the sets of size q_c for a country, as it does not determine all the preference rankings over sets. However, this does not affect this study's analysis. It is not necessary to determine the missing preference orderings over sets, and they can be in any order.

Definition (Matching). A solution to a many-to-one refugee matching problem is μ: $\mathscr{R} \bigcup \mathscr{C} \to \mathscr{R} \bigcup \mathscr{C}$, a correspondence from $\mathscr{R} \bigcup \mathscr{C}$ to $\mathscr{R} \bigcup \mathscr{C}$ such that, for every refugee family $r \in \mathscr{R}$ and participating country $c \in \mathscr{C}$:

- $\mu(r) \in \mathscr{C}$
- $\mu(c) \subseteq \mathscr{R}$ and $|\mu(c)| \leqslant q_c$
- $\mu(r) = c \Leftrightarrow r \in \mu(c)$

Let μ and ν be two matchings. A matching μ *Pareto dominates* matching ν at preference profile $P \in \mathscr{P}$ if for all refugees $r \in \mathscr{R}$, $\mu_r W_r \nu_r$, and there exists a refugee r such that $\mu_r P_r \nu_r$. A matching μ *weakly Pareto dominates* a matching ν at P if either μ Pareto dominates ν or μ is the same as ν.
The set of problems is denoted with $\mathscr{P}$, which is

$$\langle \mathscr{R} = \bigcup_{t=1}^{T} \mathscr{R}_t, \mathscr{C}, (q_c)_{c \in \mathscr{C}}, P, \succ \rangle.$$

Let $\mathscr{M}$ be the set of matchings. Since everything else, except P is fixed, I define a mechanism as follows.

Definition (Mechanism). A mechanism is a mapping that assigns a matching to each country acceptance problem $P \in \mathscr{P}$. Formally, a mechanism is a mapping $f : P \to \mathscr{M}$.

Let f and g be two mechanisms. A *mechanism f Pareto dominates mechanism g* for $\mathscr{R}' \subseteq \mathscr{R}$ if for all profiles $P \in \mathscr{P}$, $f_{\mathscr{R}'}(P)$ weakly Pareto dominates $g_{\mathscr{R}'}(P)$, and there exists a $\bar{P} \in \mathscr{P}$ such that $f_{\mathscr{R}'}(\bar{P}) \neq g_{\mathscr{R}'}(\bar{P})$. If *mechanism f weakly Pareto dominates mechanism g*, then, at every P, mechanism f produces a matching that weakly Pareto dominates the matching produced by g.

Since this study focuses on the refugee-proposing DA and its modifications, the matching results differ with respect to the type of country profile used. A refugee assignment mechanism requires refugees to submit preferences for countries and selects a matching based on these preferences and refugee priorities. Note that the notation μ^{DA} is used for matching results obtained from the refugee-proposing DA applied to P and $\succ$.

Definition (Blocking Pairs). A matching μ is blocked by a refugee–country pair $(r,c) \in \mathscr{R} \times \mathscr{C}$ if they prefer each other, relative to μ:

1. the refugee family r prefers c to the country to which it is matched μ (i.e., $c \, P_r \, \mu(r)$), and

2. given $r \notin \mu(c)$,

    a. either the country prefers r to some refugee r' that the country is matched to in μ (i.e., $r \succ_c r'$ where $r' \in \mu(c)$)

    b. or refugee r is acceptable to country c and the country has fewer refugee families assigned to it than its quota (i.e., $r \succ_c c$ and $|\mu(c)| < q_c$).

Regarding the definitions, recall that this study uses the term "refugee" when referring to a "refugee family." We now formally define stability using the notion of blocking pairs, which will be important for the axioms studied throughout this study.

Definition (Stability). A matching is stable if it is not blocked by a refugee–country pair. A mechanism is stable if it assigns a stable matching to each country acceptance problem P.

We now adapt the well-known algorithm of Gale & Shapley (1962) to our current model and call it the *Refugee-Proposing Deferred Acceptance (DA) Mechanism*.

Step 1: Each refugee proposes their first choice country. Each country tentatively assigns its refugee residency permits to its proposers based on its quota, following its preference order. Subsequently, any remaining proposers

are rejected.

In general, at

Step k: Each rejected refugee, in the previous step, proposes the succeeding country of choice. Each country considers the refugees it has been holding along with its new proposers and tentatively assigns its refugee residency permits up to its quota, following its preference order. Subsequently, any remaining proposers are rejected.

The mechanism terminates when no refugee proposal is rejected, and each refugee is assigned the final tentative assignment. We refer to this mechanism as the *Gale-Shapley refugee-optimal stable mechanism*.

Gale & Shapley (1962) call this stable mechanism *optimal* if every refugee is at least as well off under it as under any other stable matching. Furthermore, the DA procedure yields not only a stable matching but also an optimal one (Gale & Shapley, 1962). Every refugee is at least as well off under the matching assigned by the DA mechanism as they would be under any other stable matching. This holds for both sides—refugees and countries. For every country acceptance problem, there exists a refugee-optimal stable matching, which is at least as agreeable to each refugee as any other stable matching. There also exists a country-optimal stable matching, which is at least as agreeable to each country as any other stable matching. The refugee-proposing DA mechanism leads to the refugee-optimal stable mechanism. Throughout this study, DA's matching result is denoted by μ^{DA}.

4. WEAK STABILITY

When applying a mechanism using the forced priority class profile π^E there may exist blocking pairs due to how refugees are ranked according to country preferences $\succ$. The discrepancy between refugees' ranks in π^E and $\succ$ may result in a stable mechanism with respect to π^E being no longer stable with respect to $\succ$. Therefore, when a refugee family forms a blocking pair with a country, this blocking will always be with respect to another refugee family in a higher forced PC rather than the refugee family forming a blocking pair. The basis for blocking is the priority reversal resulting from the forced priority classes. This observation inspires the following axioms, the first of which is

my definition of the first weak fairness axiom of this study.

Definition (PC Fairness Axiom). A matching μ satisfies PC fairness at a particular profile P if, for every refugee–country pair (r,c) such that r prefers c to own assignment at μ and is preferred by c to a refugee $\hat{r}$ assigned to c at μ, $\hat{r}$ is in a higher PC than r. A mechanism f satisfies PC fairness if, for every profile P, it assigns a matching μ that is PC fair.

According to PC fairness, if a refugee–country blocking pair (r,c) is such that r is preferred by c to a refugee $\hat{r}$ assigned to c, then it must be the case that $\hat{r}$ is in a higher PC than r. In order to discuss the intuition behind the axiom, we can say that the PC fairness axiom tracks the discrepancy between forced priorities and country preferences, and it allows for certain salient blocking pairs accordingly. Since we are weakening stability to respect the urgency of placing refugees who are prioritized in higher classes, stability implies PC fairness. PC fairness is an axiom that requires stability within each PC of a given hierarchical structure, based on country preferences that are preserved within each PC.

Definition (PC No-Envy Axiom). A matching μ satisfies PC no-envy at a particular profile P if $\hat{r}$ is in a higher forced priority class than r and $r \in \mu(c)$, then $\mu(\hat{r})\ W_{\hat{r}}\ c$. A mechanism f satisfies PC no-envy if, for every profile P, it assigns a matching μ that satisfies PC no-envy.

A case in point: For each preference profile P, given the matching μ assigned to P where r is among the refugees matched to c, if another refugee $\hat{r}$ is in a higher forced priority class than r, then $\hat{r}$ does not have envy for r.

To demonstrate the independence of the PC no-envy and PC fairness axioms, observe that, given a fixed profile P, the matching result of the DA, specifically μ^{DA}, is fair and thus satisfies PC fairness. However, μ^{DA} does not satisfy PC no-envy, as it is not stable with respect to the forced priority classes of π^E at each profile P. Hence, PC fairness does *not* imply PC no-envy. Consider the following example to intuitively visualize this case:

Example 1. Let us consider the problem P with refugees $\mathscr{R} = \{1,2,3,4\}$, countries $\mathscr{C} = \{a,b,c,d\}$, country quotas $q_c = 1$ for all $c \in \mathscr{C}$, and forced priority classes $\mathscr{R}_1 = \{1,2\}$ and $\mathscr{R}_2 = \{3,4\}$. Allocations of the matching result μ^{DA} at the given preference profile, in this example, are in *bold*.

P_1	P_2	P_3	P_4
c	b	c	b
b	c	b	**c**
d	**d**	**a**	a
a	a	d	d

Refugee Preferences P

$\succ_a$	$\succ_b$	$\succ_c$	$\succ_d$
1	1	4	4
4	3	2	3
2	2	1	1
3	4	3	2

Country Preferences $\succ$

π_a^E	π_b^E	π_c^E	π_d^E
1	**1**	2	1
2	2	1	**2**
4	3	**4**	4
3	4	3	3

Forced Profile π^E

Applying the refugee-proposing DA algorithm to $\succ$ and P gives us the following:

$$\mu^{DA} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ b & d & a & c \end{pmatrix}$$

The matching μ^{DA} satisfies PC fairness because it satisfies fairness. However, matching μ^{DA} does not satisfy PC no-envy. This is because $(1,c)$ and $(2,c)$ are blocking pairs for μ^{DA} with respect to π^E. Refugees 1 and 2 are envious of refugee 4, who is in a lower PC than both the refugees and matched to c in μ^{DA}.

Furthermore, PC no-envy does *not* imply PC fairness. Let μ_E^{DA} be the matching result of the DA when applied to forced hierarchical priority classes. μ_E^{DA} is stable with respect to the forced hierarchical priority classes of π^E at each profile P and hence satisfies PC no-envy. Notably, observe that PC no-envy is satisfied whenever a serial dictatorship procedure is used with the permutation of agents based on the forced hierarchical priority classes. For the purpose of this exercise, suppose an arbitrary common priority ordering of agents π and let f denote a serial dictatorship procedure. Recall that, given an ordering π of agents with any permutation of the entire set of agents, a serial dictatorship $f(\pi)$ (Satterthwaite & Sonnenschein, 1981) assigns the objects to

agents as follows: The first agent is assigned their first choice among all the objects. The second agent is assigned their first choice among all the objects, excluding the choice of the first agent, and so on. Now, consider an ordering π of refugees following the fixed hierarchy of priority classes; any ordering is possible within the priority classes as long as it is the same for each country. Then, the PC no-envy property is satisfied whenever $f(\pi)$ is used since $f(\pi)$ assigns the permits to refugees following the given common order. Consider the example below to examine this visually.

Example 2. Consider the problem P below, where refugees $\mathscr{R} = \{1,2,3,4\}$, countries $\mathscr{C} = \{a,b,c,d\}$, and country quotas $q_c = 1$ for all $c \in \mathscr{C}$, and forced priority classes $\mathscr{R}_1 = \{1,2\}$ and $\mathscr{R}_2 = \{3,4\}$. Let π be the permutation. Allocations of the result of the serial dictatorship f using the given common priority order π are in bold.

P_1	P_2	P_3	P_4
a	**a**	**c**	b
b	c	b	c
c	d	a	a
d	b	d	**d**

Refugee Preferences P

π_a	π_b	π_c	π_d
2	2	2	2
1	1	1	1
3	3	3	3
4	4	4	4

Common Priority Order π

π_a^E	π_b^E	π_c^E	π_d^E
1	2	2	1
2	**1**	1	2
4	4	**3**	**4**
3	3	4	3

Forced Profile π^E

The result of the serial dictatorship $f(\pi)$ is as follows:

$$f(\pi) = \begin{pmatrix} 1 & 2 & 3 & 4 \\ b & a & c & d \end{pmatrix}$$

Observe that $f(\pi)$ satisfies PC no-envy as no refugee in a higher forced priority class envies a refugee in a lower priority class. However, it does not

satisfy PC fairness because $(1,a)$ is a blocking pair for $f(\pi)$ with respect to π^E. Refugee 1 is envious of 2, who is preferred by a over 2, according to country a's preferences within the top PC of a. Hence, country preferences are violated by $f(\pi)$ within the PC. Therefore, whenever $f(\pi)$ is applied, PC no-envy is satisfied since it is impossible to be envious of a refugee in a lower priority class. However, the result $f(\pi)$ fails PC fairness, as any ordering of refugees is possible within a priority class.

Next, let's formally define the DA that runs taking into account the hierarchical priority classes.

Definition (The DA with Hierarchical Priority Classes). The DA with hierarchical priority classes is the refugee-proposing DA applied to the forced priority profile of countries π^E and refugee preference profile P.

The matching result of the DA with hierarchical priority classes is denoted by μ_E^{DA}.

Proposition 1. Stability with respect to hierarchical priority classes is equivalent to PC no-envy and PC fairness.

Proof. "If" Part. We show that stability with respect to hierarchical priority classes implies PC no-envy and PC fairness. Let μ_E^{DA} be the matching result of the DA with hierarchical priority classes for a given problem P. Given that the DA with hierarchical priority classes is the DA applied to forced priorities π^E, then μ_E^{DA} will be stable with respect to π^E for any given preference profile. Then, there will be no refugee–country pair (r,c) blocking μ_E^{DA} for any given preference profile. This implies two cases.

Case 1. Suppose PC fairness is not satisfied. Then there is a refugee–country pair (r,c) such that r prefers c and is preferred by c for a refugee $\hat{r}$ assigned to c, where r and $\hat{r}$ are in the same forced PC. Hence, there is a ranking violation within a priority class. However, this is a direct contradiction to the stability of the DA procedure with hierarchical priority classes, which is applied with respect to forced priorities π^E.

Case 2. Suppose PC no-envy is not satisfied. Then there is a refugee–country pair (r,c) such that r prefers c and is ranked highly by c, according to c's priority ranking in π^E to a refugee $\hat{r}$ assigned to c, where r and $\hat{r}$ are not in the same forced PC. Then, there is a ranking violation across priority classes.

Specifically, then there is r who is in a higher PC than $\hat{r}$ and has envy for $\hat{r}$. However, this is a direct contradiction to the stability of the DA procedure that is applied using the forced hierarchical priority classes of π^E.

Therefore, the stability of DA with the forced hierarchical priority classes of π^E implies PC no-envy and PC fairness.

"Only If" Part. We show that satisfying PC fairness and PC no-envy implies satisfying stability with respect to hierarchical priority classes.

Case 1. Suppose PC fairness is satisfied. Then, if a refugee–country pair (r,c) is such that r prefers c and is preferred by c for a refugee $\hat{r}$ assigned to c, then $\hat{r}$ must be in a higher PC than r. This implies $\hat{r}$ and r are not in the same PC. Hence, in the same PC, there is no refugee–country pair (r,c) such that r prefers c and is preferred by c for a refugee $\hat{r}$ assigned to c. This satisfies stability within each forced PC.

Case 2. Suppose PC no-envy is satisfied. For each preference profile P, if r is in a higher forced PC than $\hat{r}$, then r does not envy $\hat{r}$. Hence, for any given preference profile P, if $\hat{r} \in \mu_E^{DA}(c)$, then $\mu_E^{DA}(r)\ W_r\ c$. Therefore, there is no refugee–country pair (r,c) such that r prefers c and is ranked above by c according to c's priority ranking in π^E to a refugee $\hat{r}$ assigned to c. This satisfies stability across forced hierarchical priority classes.

Therefore, PC fairness and PC no-envy imply stability with respect to π^E. In conclusion, satisfying the two axioms of PC fairness and PC no-envy is equivalent to satisfying stability with respect to the forced hierarchical priority classes. □

A matching is called *optimal with respect to a stability axiom* if every agent receives at least as good an assignment in this matching as in any other matching satisfying the stability axiom. Specifically, a matching μ is *optimal with respect to PC fairness and PC no-envy* at profile P if, for each refugee $r \in \mathscr{R}$, $\mu(r) = c$ is the most preferred country among all the countries that refugee r could be matched to at any matching satisfying PC fairness and PC no-envy at P, when there is any such country. Given $P \in \mathscr{P}$, a matching mechanism is *optimal with respect to PC fairness and PC no-envy* if, for each preference profile $P \in \mathscr{P}$, it assigns a matching to profile P that is optimal with respect to PC fairness and PC no-envy.

Moreover, the DA with hierarchical priority classes is the DA where countries' preferences are obtained from the ordered priority classes. Correspond-

ingly, within each class, each country c break ties according to $\succ_c$. Hence, the DA with hierarchical priority classes is a unique mechanism that is optimal with respect to PC fairness and PC no-envy.

Theorem 1. A mechanism satisfies PC fairness, PC no-envy, and is optimal with respect to PC fairness and PC no-envy if, and only if, it is the DA with hierarchical priority classes.

Proof. "If" Part. The DA with hierarchical priority classes satisfies PC fairness and PC no-envy. Therefore, this part follows directly from the equivalence between stability with respect to hierarchical priority classes and the combination of PC no-envy and PC fairness.

"Only If" Part. We show that if a mechanism is PC-fair, PC no-envy, and is optimal with respect to PC fairness and PC no-envy, then it is the DA with hierarchical priority classes. First, we know that stability with respect to π^E is equivalent to the combination of PC fairness and PC no-envy. By this equivalence result in Proposition 1, the DA with hierarchical priority classes is optimal with respect to PC fairness and PC no-envy. Focusing on the welfare of refugees, maximized welfare is obtained by the refugee-proposing DA subject to stability with respect to π^E. Hence, this follows from the DA refugee-optimal solution with the priorities π^E. Consequently, since optimality implies uniqueness, we conclude that if a mechanism is PC-fair, PC no-envy, and is optimal with respect to PC fairness and PC no-envy, then it is the DA with hierarchical priority classes. □

It is vital to discuss this result intuitively apropos of the two key weak stability and fairness axioms examined in this study. PC no-envy of any arbitrary matching mechanism implies that a refugee $\hat{r}$ in a higher forced priority class does not have envy of any of the allocations below $\hat{r}$ at any profile P. This implies that the allocation starts at the first PC and follows a serial procedure of priority classes according to a given order of refugees having the fixed hierarchy of priority classes. In this case, any order can be followed within priority classes. We allocate refugees to the top PC, followed by the second PC, and so on, which is equivalent to the serial dictatorship procedure with the permutation of agents based on the hierarchical priority classes. Succeeding the completion of the top priority class allocation, those allocations are finalized and removed. Subsequently, allocations are done for the second priority class

refugees, and so on. Instead of individual agents picking their top choices, we have priority classes of refugees who get their top choices, independent of the preferences of refugees in other priority classes.

In addition to PC no-envy, the requirements of PC fairness and optimality for a matching mechanism imply that, for any problem P, the match must be fair within each PC. Thus, the DA must be applied to each PC to ensure fair allocation within each PC. Therefore, since the DA applied to P with a given π^E is the DA with hierarchical priority classes, and π^E is a partition profile, it is equivalent to the DA applied to the top PC than the DA applied to the second PC, and so on, following the given hierarchical priority class ordering. Therefore, if a mechanism satisfies PC fairness, PC no-envy, and is optimal with respect to PC fairness and PC no-envy, then it is the DA with hierarchical priority classes.

5. TOP STABILITY

It is intuitive and vital to consider the top choice countries of refugees when designing a refugee matching system. We recognize that the imposed priority classes force the countries to change their preference rankings, deviating this study from the objective of a fair refugee allocation. Since the imposed priorities give certain refugees a priority in each country, it is crucial to incorporate refugees' preferences into the matching algorithms. A refugee in a lower PC always remains in that PC. Since priority classes are forced in all countries, every refugee in the first PC is always prioritized over every refugee in the second PC. Considering the refugees in the lower priority classes, the proposed model gives these refugees a better chance of matching with their top-ranked country by lifting them up, using their preferences. For interesting and relevant notions, which were studied independently and in different contexts, please refer to Morrill (2013); Pathak & Sönmez (2013); Afacan et al. (2017); and Dur et al. (2018).

For the definitions of weak stability axioms, recall the definition of blocking pairs (r, c) in Section 3. In addition, please note that priority classes that are common to all countries are used in every section. Sections differ only in the newly introduced mechanism designs.

Definition (Top Stability Axiom). A matching is top stable if it cannot be

blocked with a pair (r, c), where P_r ranks country c first at a fixed preference profile P. A mechanism is top stable if, for each preference profile P, whenever the matching assigned to P is blocked by (r, c), P_r does not rank c first.

Definition (Top PC Fairness Axiom). A matching μ satisfies top PC fairness if, for every refugee–country (r, c) such that r prefers c at μ and is preferred by c to a refugee $\hat{r}$ assigned to c at μ,

- either $\hat{r}$ is in a higher forced PC than r,
- or $P_{\hat{r}}$ ranks c first, and r is not in PC $\mathscr{R}_1$ of country c.

A mechanism f satisfies top PC fairness if, for every profile P, it assigns a matching μ that is top PC-fair. Notably, whenever there is justified envy with respect to $\hat{r}$, either $\hat{r}$ is in a higher forced class than r with justified envy or $P_{\hat{r}}$ ranks c first, and r is not in the top PC of country c.

To accommodate more refugees in the lower priority classes by raising them to the first PC at their top-ranked country, I further weaken the PC fairness axiom. Therefore, the PC fairness axiom implies top PC fairness. Since I make exceptions for these refugees regarding their preferred country and account for their exceptions of gained rank, I declare blocking pairs involving such refugees and their top choice country to be salient and, therefore, inadmissible. In addition, top stability and top PC fairness axioms are not independent of each other. Connections between top stability and top PC fairness axioms are demonstrated in Example 3.

Moreover, in order to obtain a combined priority ranking profile for countries that merge $\succ$ with π^E, we would need a mechanism that outlines a method to combine the two profiles $\succ$ and π^E. Such a mechanism would ideally be based on refugee preferences P when combining the two profiles $\succ$ and π^E.

Remark. The combined priority profile of countries depends on the refugee preference profile P. However, the forced priority profile of countries is independent of the refugee preference profile P.

One of the two mechanisms designed in this study is the following:

Definition (The Top Prioritization Mechanism).

1. For every refugee r, modify π_c^E by lifting refugee r up to PC $\mathscr{R}_1$ of c, where c is top-ranked in P_r.

2. Position r within PC $\mathscr{R}_1$, according to $\succ_c$ (i.e., keep all the other priority rankings the same, except for r's). This yields the new combined profile π^{E-top}.

3. Apply the DA to the π^{E-top}. This yields the matching result $f_{E-top}^{DA}(P)$ of the top prioritization mechanism f_{E-top}^{DA} at problem P.

Similar to the other notations for the matching outcomes, I use μ_{E-top}^{DA} for the matching result of the top prioritization mechanism f_{E-top}^{DA}, for any given problem P.

Furthermore, given an ordering π_c, let $S_c^{\pi}(r)$ denote the *upper contour set* at r. Then, the upper contour set of refugee r in country c's profile is as follows:

$$S_c^{\pi}(r) = \{\hat{r} \in \mathscr{R} : \hat{r}\ \pi_c\ r\ \}.$$

Lemma 1. If P_r ranks country c first, then $S_c^{\pi^{E-top}}(r) \subseteq S_c^{\succ}(r)$.

Proof. Let r and c be such that P_r ranks country c first. Then, r is lifted to $\mathscr{R}_1$ in π_c^{E-top}. Fix $\hat{r} \in S_c^{\pi^{E-top}}(r)$. Then, we have $\hat{r} \in \mathscr{R}_1$. Moreover, given the construction of π^{E-top}, this means that $\hat{r} \succ_c r$. Thus, $\hat{r} \in S_c^{\succ}(r)$. □

In addition, the top prioritization mechanism guarantees the top stability of μ_{E-top}^{DA} at any P. Thus, we have the following result:

Theorem 2. The top prioritization mechanism is

1. top stable and

2. top PC-fair.

Proof. 1. Top stability: Fix a given preference profile P. For contradiction, suppose the matching result μ_{E-top}^{DA} is not top stable at the given profile P. Then, there is a pair (r,c) blocking μ_{E-top}^{DA} with respect to $\succ$ at P, where refugee r ranks c first at preference profile P. Given that the DA is applied to the combined profile π^{E-top}, if country c rejects r in any round of the DA,

then c is temporarily matched to at least one refugee other than r. Let this refugee be $\hat{r}$. Then, this implies $\hat{r}\ \pi_c^{E-top}\ r$. Since r ranks c first at preference profile P, r is lifted up to PC $\mathscr{R}_1$ by country c in π_c^{E-top} by the top prioritization procedure. By Lemma 1, we conclude the following. If r is in the top PC of π_c^{E-top} and $\hat{r}\ \pi_c^{E-top}\ r$, then r and $\hat{r}$ will both be in the top PC. Then, $\hat{r} \succ_c r$ since country preferences are preserved within each PC. This contradicts the assumption that (r,c) is a blocking pair of μ_{E-top}^{DA} with respect to $\succ$. Hence, μ_{E-top}^{DA} is top stable.

2. Top PC-fairness: Fix a given preference profile P. For contradiction, suppose μ_{E-top}^{DA} is not top PC-fair at the given profile P. Then, at the given P, there exists a salient pair (r,c) that blocks μ_{E-top}^{DA} with respect to $\succ$ that is not allowed under top PC fairness. There is $\hat{r} \in \mathscr{R}$ such that $\hat{r} \in \mu_{E-top}^{DA}(c)$. The violation of the top PC fairness of the matching result μ_{E-top}^{DA} at the given P implies four non-trivial cases. We verify the contradictions under each case.

First, observe that when refugee $\hat{r}$ is not in a higher forced PC than r who is assumed to be blocking with c at $\succ$, then they are in the same PC. Although this is trivial, it is a possibility because the matching result is assumed to not be top PC-fair at the given profile. This violates country preferences within the same PC, which are assumed to be preserved under the top prioritization mechanism. Hence, this is a direct contradiction.

Case 1: Suppose $P_{\hat{r}}$ does not rank c first and r is in PC $\mathscr{R}_1$ of country c. Then, either $\hat{r}$ is in a lower PC than r, or $\hat{r}$ is already in $\mathscr{R}_1$. Hence, both r and $\hat{r}$ are in $\mathscr{R}_1$ of country c. If $r \in \mathscr{R}_1$ of c and $\hat{r}$ is in a lower PC, then this contradicts $\hat{r}\ \pi_c^{E-top}\ r$ since $\hat{r} \in \mu_{E-top}^{DA}(c)$. If both $r, \hat{r} \in \mathscr{R}_1$, then it must be that $r \succ_c \hat{r}$, since for contradiction we assumed the existence of a blocking pair (r,c) with respect to $\succ$. Country preferences are preserved within the same priority classes. Hence, $r\ \pi_c^{E-top}\ \hat{r}$. However, we also have $\hat{r}\ \pi_c^{E-top}\ r$ since $\hat{r} \in \mu_{E-top}^{DA}(c)$. Thus, this is a contradiction.

Case 2: Suppose $P_{\hat{r}}$ ranks c first and r is in PC $\mathscr{R}_1$ of country c. Then, $r, \hat{r} \in \mathscr{R}_1$, then $r \succ_c \hat{r}$, which is preserved in $\mathscr{R}_1$ and hence, $r\ \pi_c^{E-top}\ \hat{r}$. However, this contradicts $\hat{r}\ \pi_c^{E-top}\ r$ given $\hat{r} \in \mu_{E-top}^{DA}(c)$.

Case 3: Suppose $P_{\hat{r}}$ does not rank c first and r is not in PC $\mathscr{R}_1$ of country c. If $\hat{r}$ is in forced class $\mathscr{R}_1$, then $\hat{r}\ \pi_c^{E-top}\ r$, and so there is nothing to prove as μ_{E-top}^{DA} is top PC-fair. However, consider otherwise. Then, neither of the refugee families is in $\mathscr{R}_1$. Since we assumed μ_{E-top}^{DA} is not top PC-fair, either

r and $\hat{r}$ are in the same PC or $\hat{r}$ is in a lower PC than r. If they are both in the same PC, then (r,c) blocking μ^{DA}_{E-top} at $\succ$ will imply that $r \succ_c \hat{r}$, which is preserved in the same PC, then $r\ \pi^{E-top}_c\ \hat{r}$. This is a contradiction to $\hat{r}\ \pi^{E-top}_c\ r$, given $\hat{r} \in \mu^{DA}_{E-top}(c)$. If $\hat{r}$ is in a lower PC than r, then r having envy for $\hat{r}$ in a lower PC will be a contradiction of the PC no-envy property of the match. This is because if r is in a higher PC than $\hat{r}$, then r cannot have envy for $\hat{r}$ at preference profile P.

Case 4: Suppose $P_{\hat{r}}$ ranks c first and r is not in PC $\mathscr{R}_1$ of country c. Then, (r,c) is a salient blocking pair that is allowed under top PC fairness. Hence, this is a direct contradiction.

Therefore, the top prioritization mechanism is top stable and top PC-fair. □

The next example demonstrates how we achieve the combination of the forced priority classes and the preference rankings of countries, and how we further modify the resulting country rankings by using the top prioritization mechanism to provide higher rank for refugees with their top-ranked countries. The example also verifies that the resulting matching satisfies the weakened axioms of top stability and top PC fairness.

Example 3. Consider the given refugee families $\mathscr{R} = \{1,2,3,4,5,6\}$, countries $\mathscr{C} = \{a,b,c,d\}$, quota vector $q = (1,1,2,2)$, and forced priority classes $\mathscr{R}_1 = \{1,2\}$ and $\mathscr{R}_2 = \{3,4,5,6\}$. Since countries c and d have two residency quotas each, let there be two identical copies of country c, each with preferences $\succ_c$ and capacity one. Similarly, let there be two identical copies of d each with $\succ_d$ and capacity one. The underlined allocations are for the matching result μ^{DA}_{E-top} of the top prioritization mechanism.

P_1	P_2	P_3	P_4	P_5	P_6
d	d	$\underline{d}$	$\underline{d}$	a	a
$\underline{a}$	$\underline{b}$	c	c	d	d
b	c	a	a	$\underline{c}$	$\underline{c}$
c	a	b	b	b	b

Refugee Preferences P

$\succ_a$	$\succ_b$	$\succ_c$	$\succ_d$
$\underline{1}$	$\underline{2}$	3	5
$\underline{5}$	$\underline{1}$	4	6
6	5	$\underline{5}$	$\underline{3}$
2	6	$\underline{6}$	$\underline{4}$
3	3	$\underline{2}$	$\underline{1}$
4	4	1	2

Country Preferences $\succ$

π_a^E	π_b^E	π_c^E	π_d^E
1	2	2	1
2	1	1	2
5	5	3	5
6	6	4	6
3	3	5	3
4	4	6	4

Forced Profile π^E

π_a^{E-top}	π_b^{E-top}	π_c^{E-top}	π_d^{E-top}
1	2	2	3
5	1	1	4
6	5	3	1
2	6	4	2
3	3	5	5
4	4	6	6

Combined Profile π^{E-top}

Observe that P_1, P_2, P_3, and P_4 rank country d first, and P_5 and P_6 rank country a first. Notably, countries b and c are not popular enough among refugees to be top-ranked. Thus, the following countries lift to their top PC $\mathscr{R}_1$ those refugees that top-ranked them: country d lifts up $\{1,2,3,4\}$ and country a lifts up 5 and 6. This gives us the combined priority profile π^{E-top} above.

After applying the DA to π^{E-top} and P, we obtain the matching below:

$$\mu_{E-top}^{DA} = \begin{pmatrix} a & b & c & d \\ 1 & 2 & \{5,6\} & \{3,4\} \end{pmatrix}$$

Observe that the blocking pair set of μ_{E-top}^{DA} with respect to $\succ$ comprises only $(5,d)$ and $(6,d)$. Since d is not the top choice of either of the refugees 5 and 6, the top stability of μ_{E-top}^{DA} is satisfied for the given problem P. Moreover, the reason for refugees 5 and 6 blocking μ_{E-top}^{DA} with country d with respect to $\succ$ is the loss of country d to refugees 3 and 4. It is also attributed to the priority reversal between $\{5,6\}$ and $\{3,4\}$, which stems from P_3 and P_4 ranking d as their top choice, and this leads to $\{3,4\}$ getting lifted to the top priority class of country d and gaining a rank over refugees 5 and 6. Therefore, top PC fairness of μ_{E-top}^{DA} is satisfied.

In this example, we can also see how top stability can imply top PC fairness. In the combined profile, π^{E-top} refugees are already moved up to the top PC of their top choice countries. Refugees $\{5,6\}$ have justified envy of $\{3,4\}$ since $\{3,4\}$ are moved up to the top PC of d. The blocking pairs $(5,d)$ and $(6,d)$ are therefore allowed under top PC fairness.

A mechanism f is *strategy-proof* if for all $r \in \mathscr{R}$, all $P \in \mathscr{P}$, and all P_r', $f_r(P)\ W_r\ f_r(P_r', P_{-r})$. A coalition $L \subseteq \mathscr{R}$ can *manipulate* matching $f(P)$ at P

if there exists P'_L such that for all $r \in L$, $f_r(P'_L, P_{-L})\ P_r\ f_r(P)$.[4]

Remark. The top prioritization mechanism is not strategy-proof for refugee families. Now we will prove this statement. Let P be the original refugee preference profile and P' be the misreported refugee preference profile. Let $P_4 = aP_4dP_4cP_4b$ such that the only difference between these two profiles is that refugee 4 misreports its own top choice country as d instead of the truthful a. For simplicity, suppose $q_c = 1$ for each $c \in \mathscr{C}$. It must be noted that refugee $f^{DA}_{E-top}(P)(4) = c$, where $f^{DA}_{E-top}(P)(4)$ is the outcome assigned to refugee 4 by the top prioritization mechanism f^{DA}_{E-top} at P. Regard the given priority classes $\mathscr{R}_1 = \{1,2\}$ and $\mathscr{R}_2 = \{3,4\}$.

P_1	P_2	P_3	P'_4
d	d	d	d
a	b	c	a
b	c	a	c
c	a	b	b

Refugee Preferences P'

$\succ_a$	$\succ_b$	$\succ_c$	$\succ_d$
1	2	3	4
4	1	4	3
2	4	2	1
3	3	1	2

Country Preferences $\succ$

π^E_a	π^E_b	π^E_c	π^E_d
1	2	2	1
2	1	1	2
4	4	3	4
3	3	4	3

Forced Profile π^E

$\pi^{E-top'}_a$	$\pi^{E-top'}_b$	$\pi^{E-top'}_c$	$\pi^{E-top'}_d$
1	2	2	4
2	1	1	3
4	4	3	1
3	3	4	2

Combined Priority $\pi^{E-top'}$

After applying the DA to $\pi^{E-top'}$ and P', we have the matching below:

$$\mu^{DA'}_{E-top} = \begin{pmatrix} a & b & c & d \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

Therefore, refugee 4 manipulates the top prioritization mechanism f^{DA}_{E-top} at P since there exists P'_4 such that $f^{DA}_{E-top}(P'_4, P_{-4})(4)\ P_4\ f^{DA}_{E-top}(P)(4)$. Hence, refugee 4 manipulates the mechanism by getting their untruthful top choice.

[4] Alternative notations to $f_r(P'_L, P_{-L})$ and $f_r(P)$ are $f(P'_L, P_{-L})(r)$ and $f(P)(r)$, respectively.

6. CREDIBLE STABILITY

Turning to the refugees forced to be in the lower priority classes, I recognize the need for and importance of exploring the means of giving these refugees an additional chance of attaining a higher rank. As mentioned, I explored this by considering two different forms of priority profiles. The first form is obtained by the top prioritization mechanism, which gives these refugees a better chance of matching with their top-ranked country. In this section, I describe my second form, under which I provide these refugees a better opportunity of accessing their DA-matched country. For a relevant mechanism studied independently and in a school choice context, please see Biró & Gudmundsson (2021). In their study, they investigate the complexity of finding Pareto-efficient allocations of highest welfare by providing top priority to all students at the schools where they are assigned in the socially optimal solution.

Unlike the top prioritization mechanism, I find that the DA matching is the best for refugees in a stable matching, which shows the importance of prioritizing these refugees in their DA-matched countries. This prioritization would also be more compelling for countries because *each* country's priorities must undergo *fewer* quota-based modifications than the top prioritization mechanism. This is because, under the top prioritization mechanism, it will not be necessary for some countries to modify their priority orders as much as the popular countries. For example, less-preferred countries for settlement, such as Poland, are less likely to be top-ranked by refugees. These countries will have fewer refugees requiring a promotion to the top PC. However, popular countries, such as Germany, may have to modify their priority rankings to a greater extent. If all refugees rank Germany as their top country of choice, then all refugees will be moved up to the top PC of Germany. Within a class, Germany will rank according to its own preferences based on its points system. If a country is popular, it will be required to make several modifications to its priority ordering, thereby deviating more from the forced priority classes.

I start by defining a *credible blocking pair* to build the new weak stability axiom. Here, I would like to remind the readers that the matching result of the DA ran with $\succ$, and P is denoted by μ^{DA}.

Definition (Credible Blocking Pair). A pair blocking a matching μ is credible if it is matched under μ^{DA}.

A pair blocking a matching μ is *non-credible* if it is *not* matched under μ^{DA}.

Definition (Credible Stability Axiom). A matching is credibly stable if it cannot be blocked with a credible blocking pair at a given P. A mechanism is credibly stable if, for each preference profile P, whenever the matching assigned to P is blocked by (r,c), (r,c) is not a credible blocking pair at P.

Definition (Credible PC Fairness Axiom). A matching μ satisfies credible PC fairness at a particular profile P if, for every refugee–country pair (r,c) such that r prefers c at μ and is preferred by c to a refugee $\hat{r}$ assigned to c at μ,

- either $\hat{r}$ is in a higher forced PC than r,
- or $\hat{r} \in \mu^{DA}(c)$ at P and r is not in PC $\mathscr{R}_1$ of country c.

A mechanism f satisfies credible PC fairness if, for every profile P, it assigns a matching μ that is credibly PC-fair. Notably, whenever there is justified envy with respect to $\hat{r}$, either $\hat{r}$ is in a higher forced class than r with justified envy or $\hat{r} \in \mu^{DA}(c)$ at P and r is not in top PC of country c. To accommodate more refugees in lower priority classes by raising them to the first priority class at their DA-matched country, I further weaken the PC fairness axiom. Therefore, the PC fairness axiom implies credible PC fairness. Since I make exceptions for these refugees with their DA-matched country and account for their exceptions of gained rank, I declare blocking pairs that involve such refugees and their DA-matched country to be salient and inadmissible. In addition, these two axioms are not independent of each other. Connections between credible stability and credible PC fairness axioms are demonstrated as part of Example 4.

Definition (The DA-Match Prioritization Mechanism).

1. For every refugee r, modify π_c^E by lifting refugee r up to PC $\mathscr{R}_1$ of c, where $r \in \mu^{DA}(c)$.
2. Position r within PC $\mathscr{R}_1$ according to $\succ_c$ (i.e., keep all the other priority rankings the same, except for r's). This yields the new combined profile π^{E-DA}.

3. Apply the DA to π^{E-DA}. This yields the matching result $f^{DA}_{E-DA}(P)$ of the DA-match prioritization mechanism f^{DA}_{E-DA} at problem P.

Similar to the other aforementioned notations for the matching outcomes, I use μ^{DA}_{E-DA} for the matching result of the DA-match prioritization mechanism f^{DA}_{E-DA} for any given problem P.

Unlike the top prioritization mechanism, each country's DA-matched refugees get lifted to the PC $\mathscr{R}_1$. Under the top prioritization mechanism, only the top-ranked countries lift refugees up to their PC $\mathscr{R}_1$. However, I now lift the DA-matched refugees to their DA-matched country's PC $\mathscr{R}_1$. The key distinction is that the DA-match is used to promote the refugees, which is an assignment, unlike the top choices. Thus, each country gets to lift its DA-matched refugees.

Lemma 2. If $r \in \mu^{DA}(c)$, then $S_c^{\pi^{E-DA}}(r) \subseteq S_c^{\succ}(r)$.

Proof. Let r and c be such that $r \in \mu^{DA}(c)$. Then, r is lifted to $\mathscr{R}_1$ in π_c^{E-DA}. Fix $\hat{r} \in S_c^{\pi^{E-DA}}(r)$. Then, we have $\hat{r} \in \mathscr{R}_1$. Moreover, given the construction of π^{E-DA}, this means that $\hat{r} \succ_c r$. Thus, $\hat{r} \in S_c^{\succ}(r)$. □

Lemma 3. μ^{DA} is stable with respect to the priority profile π^{E-DA}.

Proof. For contradiction, suppose there is a pair (r,c) blocking μ^{DA} with respect to π^{E-DA}. Let r be one of the $|\mu^{DA}(\bar{c})|$ refugees matched to $\bar{c}$. In other words, let $r \in \mu^{DA}(\bar{c})$. Then, $c\ P_r\ \bar{c}$. Let $\hat{r}$ be one of the $|\mu^{DA}(c)|$ refugees matched to c. In other words, let $\hat{r} \in \mu^{DA}(c)$. Then, $r\ \pi_c^{E-DA}\ \hat{r}$. Thus, $r \in S_c^{\pi^{E-DA}}(\hat{r})$, and Lemma 2 implies that $r \in S_c^{\succ}(\hat{r})$. Then (r,c) is also blocking μ^{DA} with respect to $\succ$, which is a contradiction, since μ^{DA} is stable at $\succ$. □

Lemma 3 can occur in two cases: μ^{DA} is stable with respect to π^{E-DA} either by simply being equal to μ^{DA}_{E-DA} or when these two matches differ. Through the DA-match prioritization mechanism, we have $\mu^{DA}_{E-DA} = \mu^{DA}$ implying, μ^{DA} is still stable with respect to π^{E-DA}, and even μ^{DA}_{E-DA} is stable with respect to $\succ$. We can also have $\mu^{DA}_{E-DA} \neq \mu^{DA}$, where μ^{DA} is still stable with respect to π^{E-DA}, and μ^{DA}_{E-DA} Pareto dominates μ^{DA} for refugees, which I prove in the next theorem.

Remark. When μ^{DA}_{E-DA} coincides with μ^{DA}, then both μ^{DA} and μ^{DA}_{E-DA} are stable with respect to both $\succ$ and π^{E-DA}.

Definition (Weak Pareto Domination). A mechanism f weakly Pareto dominates another mechanism g if, for every P, the matching assigned to f weakly Pareto dominates the matching assigned to g.

Theorem 3. The DA-match prioritization mechanism weakly Pareto dominates the DA for the refugees.

Whenever the matching results do not coincide at a particular profile P, then the DA-match prioritization mechanism leads to a matching μ^{DA}_{E-DA} that Pareto dominates μ^{DA} for refugees.

Proof. From Gale & Shapley (1962)'s optimality result, we know that the refugee-proposing DA-match prioritization leads to the refugee-optimal stable matching with respect to π^{E-DA}. Since μ^{DA}_{E-DA} is the refugee-optimal stable matching at π^{E-DA} and matching μ^{DA} is also stable with respect to π^{E-DA} by Lemma 3, if $\mu^{DA}_{E-DA} \neq \mu^{DA}$, then μ^{DA}_{E-DA} Pareto dominates μ^{DA} with respect to P. □

Furthermore, according to Knuth (1976)'s polarity result, for every problem with strict preferences, both sides of the market have common preferences and these common preferences are opposed to each other on the set of stable matchings. We know that the refugees' preferences P and the combined profile for the countries π^{E-DA} are opposed to each other on the set of stable matchings. For example, following Knuth's polarity, consider μ and μ' as two stable matchings. Then, all the refugees like μ at least as well as μ' if and only if all countries like μ' at least as well as μ. Intuitively, the best stable matching for one side is the worst stable matching for the other side. Thus, the refugee-optimal stable matching is the worst stable matching for countries (country-pessimal), and the country-optimal stable matching is the worst stable matching for refugees (refugee-pessimal). Therefore, as per Knuth (1976) polarity result and by Lemma 3, I observe that when μ^{DA}_{E-DA} differs from μ^{DA}, we have the matching μ^{DA}_{E-DA}—the country-pessimal stable matching with respect to π^{E-DA}. Thus, from the viewpoint of countries, μ^{DA}, which is stable at π^{E-DA}, Pareto dominates μ^{DA}_{E-DA} at π^{E-DA}.

Moreover, the DA-match prioritization mechanism leads to a matching μ^{DA}_{E-DA} that satisfies credible stability. Recall that a credible blocking pair is a pair that is matched under μ^{DA}.

Theorem 4. The DA-match prioritization Mechanism is

1. credibly stable and
2. credibly PC fair.

Proof. 1. Credible stability: Fix a given preference profile P. Suppose $\mu^{DA}_{E-DA} \neq \mu^{DA}$. For contradiction, suppose the matching result μ^{DA}_{E-DA} is not credibly stable at the given profile P. Then, there is a pair (r,c) blocking μ^{DA}_{E-DA} with respect to $\succ$ that is matched under μ^{DA}. Let r be one of the $|\mu^{DA}_{E-DA}(\bar{c})|$ refugees matched to $\bar{c}$ under the DA-match prioritization mechanism. In other words, $r \in \mu^{DA}_{E-DA}(\bar{c})$. We know $\mu^{DA}_{E-DA} \neq \mu^{DA}$ and $c \neq \bar{c}$. Thus, by Theorem 3, $\bar{c}\ P_r\ c$, this contradicts the assumption that (r,c) is a blocking pair of the matching μ^{DA}_{E-DA} with respect to $\succ$, given the profile P.

2. Credible PC fairness: Fix a given preference profile P. For contradiction, suppose μ^{DA}_{E-DA} does not satisfy the credible PC fairness at P. Then, at the given P, there exists a pair (r,c) blocking μ^{DA}_{E-DA} with respect to $\succ$ that is not allowed under credible PC fairness and there exists $\hat{r} \in \mathscr{R}$ such that $\hat{r} \in \mu^{DA}_{E-DA}(c)$. The violation of the credible PC fairness of the matching result μ^{DA}_{E-DA} at the given P implies four non-trivial cases. We verify the contradiction in each case.

First, observe that when refugee $\hat{r}$ is not in a higher forced PC than r, who is assumed to be blocking with c at $\succ$, then they are in the same PC. Although this is trivial, it is a possibility since the matching result is assumed to be not credible PC-fair at the given profile. This violates the country preferences within the same PC, which are assumed to be preserved under the DA-match prioritization mechanism. Hence, this is a direct contradiction.

Case 1: Suppose $\hat{r} \notin \mu^{DA}(c)$ at given P, and r is in top PC $\mathscr{R}_1$ of country c. Then, either $\hat{r}$ is in a lower PC than r or $\hat{r}$ is already in top PC $\mathscr{R}_1$, and hence they are both in $\mathscr{R}_1$ of country c. If $r \in \mathscr{R}_1$ of c and $\hat{r}$ is in a lower PC, then this will contradict $\hat{r}\ \pi^{E-DA}_c\ r$ since $\hat{r} \in \mu^{DA}_{E-DA}(c)$. If both $r, \hat{r} \in \mathscr{R}_1$, then $r \succ_c \hat{r}$, and thus it should be preserved in $\mathscr{R}_1$. Then, $r\ \pi^{E-DA}_c\ \hat{r}$. However, this contradicts $\hat{r}\ \pi^{E-DA}_c\ r$, given $\hat{r} \in \mu^{DA}_{E-DA}(c)$.

Case 2: Suppose $\hat{r} \in \mu^{DA}(c)$ at P, and r is in PC $\mathscr{R}_1$ of country c. Then, $r, \hat{r} \in \mathscr{R}_1$ and by our assumption we have $r \succ_c \hat{r}$, which is preserved in $\mathscr{R}_1$ and so $r \; \pi_c^{E-DA} \; \hat{r}$. However, given $\hat{r} \in \mu^{DA}_{E-DA}(c)$ we also have $\hat{r} \; \pi_c^{E-DA} \; r$, which gives us a contradiction.

Case 3: Suppose $\hat{r} \notin \mu^{DA}(c)$ at P, and r is not in PC $\mathscr{R}_1$ of country c. If $\hat{r}$ is already in forced class $\mathscr{R}_1$, then $\hat{r} \; \pi_c^{E-DA} \; r$, and thus there is nothing to prove as μ^{DA}_{E-DA} is credibly PC-fair. However, considering otherwise, neither of the refugee families is in $\mathscr{R}_1$. Hence, since we assumed μ^{DA}_{E-DA} is not credible PC-fair, either r and $\hat{r}$ are in the same priority class, or $\hat{r}$ is in a lower PC than r. If they are both in the same PC, then (r,c) blocking μ^{DA}_{E-DA} at $\succ$ implies that $r \succ_c \hat{r}$, which is preserved in the same PC. Hence, $r \; \pi_c^{E-DA} \; \hat{r}$. This is a contradiction to $\hat{r} \; \pi_c^{E-DA} \; r$, given $\hat{r} \in \mu^{DA}_{E-DA}(c)$. If $\hat{r}$ is in a lower PC than r, then r having envy for $\hat{r}$ who is in a lower PC, is a contradiction to the PC no-envy property of the match. This is because if r is in a higher PC than $\hat{r}$, then r does not have envy for $\hat{r}$ at P.

Case 4: $\hat{r} \in \mu^{DA}(c)$ at P and r is not in PC $\mathscr{R}_1$ of country c. Then, (r,c) is a salient blocking pair allowed under credible PC fairness. Hence, this is a direct contradiction.

Therefore, the DA-match prioritization mechanism satisfies credible stability and credible PC fairness. □

The next example demonstrates the DA-match prioritization mechanism. It shows how we make exceptions for refugees by moving them up to the top priority class of their DA-matched countries. We also observe how the DA-match prioritization mechanism differs from the top prioritization mechanism. In contrast to the top prioritization mechanism, as seen in Example 3, we first need to obtain the DA matching result using refugee and country preferences before adjusting the country rankings accordingly. The example also shows how to verify that the weakened axioms of credible stability and credible PC fairness hold for the matching obtained using the DA-match prioritization mechanism. Finally, it demonstrates how the result of the DA-match prioritization mechanism can Pareto dominate the DA outcome for refugees.

Example 4. Consider the given refugees $\mathscr{R} = \{1,2,3,4,5,6,7,8\}$, countries $\mathscr{C} = \{a,b,c,d\}$, and forced priority classes $\mathscr{R}_1 = \{1,2,5,6\}$ and $\mathscr{R}_2 = \{3,4,7,8\}$. This example demonstrates the case where the DA-match prioritization mechanism leads to a matching μ^{DA}_{E-DA} that does not coincide with μ^{DA}.

Suppose $q = (2,2,2,2)$. The double-underlined allocations are for μ^{DA}_{E-DA}, which is obtained from applying the DA to the P and π^{E-DA}. The underlined allocations are for μ^{DA}. Allocations that are marked in bold show the instance when μ^{DA}_{E-DA} coincides with μ^{DA}.

P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
$\underline{\underline{\text{d}}}$	d	d	$\underline{\underline{\text{a}}}$	d	d	d	a
$\underline{\text{a}}$	**b**	**c**	$\underline{\text{d}}$	**a**	**b**	**c**	**d**
b	c	a	c	b	c	a	c
c	a	b	b	c	a	b	b

Refugee Preferences P

$\succ_a$	$\succ_b$	$\succ_c$	$\succ_d$
1	**2**	**3**	4
$\underline{\underline{\mathbf{5}}}$	**6**	**7**	$\underline{\underline{\mathbf{8}}}$
$\underline{4}$	1	2	3
$\underline{\underline{8}}$	5	6	7
2	3	1	$\underline{1}$
6	7	5	$\underline{\underline{2}}$
3	4	4	5
7	8	8	6

Country Preferences $\succ$

π^E_a	π^E_b	π^E_c	π^E_d
1	2	2	1
5	6	6	2
2	1	1	5
6	5	5	6
4	3	3	4
8	7	7	8
3	4	4	3
7	8	8	7

Forced Profile π^E

π^{E-DA}_a	π^{E-DA}_b	π^{E-DA}_c	π^{E-DA}_d
1	2	3	4
5	6	7	8
2	1	2	1
6	5	6	2
4	3	1	5
8	7	5	6
3	4	4	3
7	8	8	7

Combined Profile π^{E-DA}

After applying the DA to π^{E-DA} and P, we get:

$$\mu^{DA}_{E-DA} = \begin{pmatrix} a & b & c & d \\ \{4,5\} & \{2,6\} & \{3,7\} & \{1,8\} \end{pmatrix}$$

$$\mu^{DA} = \begin{pmatrix} a & b & c & d \\ \{1,5\} & \{2,6\} & \{3,7\} & \{4,8\} \end{pmatrix}$$

Applying the DA to P and π^{E-DA} gives the refugee-optimal stable matching μ^{DA}_{E-DA} at π^{E-DA}. The two matching results above are different from each other. Looking at P, observe that μ^{DA}_{E-DA} Pareto dominates μ^{DA} for refugees. At P, refugees 1 and 4 are better off under the matching μ^{DA}_{E-DA}. It must be noted that $(3,d)$ and $(7,d)$ are the only pairs blocking μ^{DA}_{E-DA} with respect to $\succ$. Since $(3,d)$ and $(7,d)$ are not matched in μ^{DA}, they are not credible. Therefore, these blocking pairs are allowed to exist under the weak stability of credible stability. Thus, the matching μ^{DA}_{E-DA} is credibly stable. In addition, the reason for refugees 3 and 7 blocking μ^{DA}_{E-DA} with respect to $\succ$ is the loss of country d to refugee 1 owing to the PC reversals between 1 and 3 and between 1 and 7. These priority reversals stem from the discrepancy between π^E and $\succ$. In other words, 1 is in a higher forced PC than 3 and 7 in π^E. This leads to the following reversal: $3 \succ_d 1$ versus $1\ \pi^E_d\ 3$ and $7 \succ_d 1$ versus $1\ \pi^E_d\ 7$. Therefore, μ^{DA}_{E-DA} *is credibly PC-fair because it is PC-fair.*

This example also shows how credible stability can imply credible PC fairness. Notably, none of the blocking pairs $(3,d)$ and $(7,d)$ are matched under the DA. In the combined profile π^{E-DA}, refugees are already lifted to the top priority class of their DA-matched countries. Therefore, the priority reversal between refugees $\{3,7\}$ and 1 is attributed to the forced priority classes—the difference between the forced profile of countries π^E and country preferences $\succ$. Refugees 3 and 7 have justified envy for 1 because 1 is forced above these other two refugees in π^E. Subsequently, these blocking pairs are allowed under PC fairness, and PC fairness implies credible PC fairness.

Remark. The DA-match prioritization mechanism is not strategy-proof for refugee families.

Proof. The matching assigned by the DA-match prioritization mechanism Pareto dominates the matching made by the DA at each preference profile. Thus, whether by Kesten (2006) and Ergin (2002), or Kesten & Kurino (2019), the DA-match prioritization mechanism cannot be strategy-proof.[5] □

[5] Kesten (2006) or Kesten & Kurino (2019) are more relevant here, given that there are no outside options for refugees in the present setting. Also note that this holds per Abdulkadiroğlu et al. (2009) too, as well as by Alva & Manjunath (2019), which require refugees to have an outside option.

7. CONCLUSION

To coordinate the stream of refugees effectively, it would be ideal to have as many European countries as possible participating in a centralized matching mechanism. Persuading all the European countries to participate in a centralized computerized matching mechanism has its challenges. This study hopes to contribute to the efforts to overcome the political impasse that tends to prevent European countries from participating in accountability-sharing in the international refugee crisis. By defining and investigating a country acceptance problem, this study proposes two matching mechanisms. These mechanisms are based on the proposed priority class hierarchy using the UNHCR humanitarian principles and guidelines. They contribute to the literature by offering methodologies that reconcile country preferences and the UNHCR-mandated hierarchical priority classes while maintaining the laudable stability and fairness properties.

Having two kinds of ranking profiles for countries, the UNHCR-mandated priority profile and the preference profile, allows for an examination of the real-world predicaments associated with the refugee reallocation problem. I capture how the difference between the two ranking profiles creates blocking pairs of countries and refugees owing to the forced hierarchical priority classes. I weaken the stability axiom since a mechanism that is stable with respect to a forced profile may no longer be stable with respect to countries' preferences. This may lead the country and refugee pairs to block a matching result. With regard to the top prioritization mechanism, I provide an additional chance to refugees forced in lower priority classes and who therefore face the risk of remaining in the lower priority classes. To this end, I prioritize these refugees in their top choice countries. The DA-match prioritization mechanism allows lifting refugees forced in lower priority classes to the first PC of their DA-matched country. I find that, for refugees, the DA-match prioritization mechanism weakly Pareto dominates the DA mechanism. Furthermore, I recognize the importance of persuading countries to participate in a centralized refugee matching mechanism. I believe that a priority class-based approach with no imposed category-specific set-aside reserve quotas may be more attractive in terms of increasing countries' willingness and incentive to participate in solving the refugee matching problem. Beyond the refugee reallocation context, this study's results have other policy applications, such as centralized college admissions, the design of public-school choice systems,

and forced migration or displacement.

This study's results may also apply in a school choice or college admission context, more so than in the context of refugee settlement. This criticism may arise because refugee allocation has some features that do not necessarily fit into the model presented. Four such factors are as follows: (i) the preferences of the refugees may not be considered or even if considered, the optimization may be focused on other factors; (ii) the size of the families can matter, as the quotas are typically for a specific number of people admitted, and further constraints may also be implied for other characteristics of the families; (iii) refugee allocation has a dynamic nature; and (iv) the usage of stability for refugee allocation in Europe may cause skepticism for political reasons, as it could cause unbalanced solutions, with the most attractive countries receiving the "best" (highly qualified, easy to settle) refugees.

With my study, I endeavour to contribute to the literature at the intersection of matching theory and refugee studies, which are less extensive than the school choice literature. This study does not seek to address all aspects of refugee matching. Addressing factors such as family size, the dynamic aspect, and finding a more balanced (fair) solution all at once can be significantly challenging. Instead, this study aims to introduce a new aspect to the choice literature, that is, hierarchical classes, which is at the proposed model's core. By emphasizing the PC hierarchy, this study highlights different theoretical aspects. Moreover, country preferences are frowned upon in refugee settlement due to discrimination issues. Therefore, hierarchical classes are more critical in the refugee settlement context as they significantly curb potential discrimination by imposing the UNHCR-based priority classes. This, therefore, makes country preferences much less important for the matching outcome.

Further, as the Syrian war prompted me to conduct this study, I focus on the static matching problem and earmark dynamic management for future work, in which I will extend my model with hierarchical classes to a dynamic setting. I will also consider investigating the incorporation of hierarchical classes in dynamic capacity management. In addition, future work should consider improving refugee allocation and integration through data-driven algorithmic assignments with hierarchical classes.

References

Abdulkadiroğlu, A., Pathak, P. A., & Roth, A. E. (2009). Strategy-proofness versus efficiency in matching with indifferences: Redesigning the NYC high school match. *American Economic Review*, *99*(5), 1954–1978.

Abdulkadiroğlu, A., & Sönmez, T. (2003). School choice: A mechanism design approach. *American Economic Review*, *93*(3), 729–747.

Afacan, M. O., Alioğulları, Z. H., & Barlo, M. (2017). Sticky matching in school choice. *Economic Theory*, *64*(3), 509–538.

Alfred, C. (2015). What history can teach us about the worst refugee crisis since WWII. *The Huffington Post*, *12*.

Alva, S., & Manjunath, V. (2019). Strategy-proof Pareto-improvement. *Journal of Economic Theory*, *181*, 121–142.

Andersson, T., & Ehlers, L. (2020). Assigning refugees to landlords in Sweden: Stable maximum matchings. *Scandinavian Journal of Economics*, *122*(3), 937–965.

Assembly, U. G., et al. (1951). Convention relating to the status of refugees. *United Nations, Treaty Series*, *189*(1), 137.

Bansak, K., Ferwerda, J., Hainmueller, J., Dillon, A., Hangartner, D., Lawrence, D., & Weinstein, J. (2018). Improving refugee integration through data-driven algorithmic assignment. *Science*, *359*(6373), 325–329.

Biró, P., & Gudmundsson, J. (2021). Complexity of finding Pareto-efficient allocations of highest welfare. *European Journal of Operational Research*, *291*(2), 614–628.

Delacrétaz, D., Kominers, S. D., Teytelboym, A., et al. (2019). *Matching mechanisms for refugee resettlement* (Tech. Rep.).

Dur, U., Mennle, T., & Seuken, S. (2018). First-choice maximal and first-choice stable school choice mechanisms. In *Proceedings of the 2018 ACM conference on Economics and Computation* (pp. 251–268).

Ergin, H. I. (2002). Efficient resource allocation on the basis of priorities. *Econometrica*, *70*(6), 2489–2497.

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

Giuffre, M., & Costello, C. (2015). Tragedy and responsibility in the Mediterranean. *Open Democracy*.

Goto, M., Iwasaki, A., Kawasaki, Y., Kurata, R., Yasuda, Y., & Yokoo, M. (2016). Strategyproof matching with regional minimum and maximum quotas. *Artificial Intelligence*, *235*, 40–57.

Jones, W., & Teytelboym, A. (2016). Choices, preferences and priorities in a match-

ing system for refugees. *Forced Migration Review*, *51*(2), 80–82.

Jones, W., & Teytelboym, A. (2017). The international refugee match: A system that respects refugee preferences and the priorities of states. *Refugee Survey Quarterly*, *36*(2), 84–109.

Kesten, O. (2006). On two competing mechanisms for priority-based allocation problems. *Journal of Economic Theory*, *127*(1), 155–171.

Kesten, O., & Kurino, M. (2019). Strategy-proof improvements upon deferred acceptance: A maximal domain for possibility. *Games and Economic Behavior*, *117*, 120–143.

Knuth, D. E. (1976). *Mariages stables et leurs relations avec d'autres problemes combinatoires: introduction a l'analysis mathematique des algorithmes-*. Les Presses de l'Universite de Montreal.

Koser, K. (2011). *Responding to migration from complex humanitarian emergencies: Lessons learned from libya*. Chatham House London.

Morrill, T. (2013). An alternative characterization of top trading cycles. *Economic Theory*, *54*(1), 181–197.

Pápai, S. (2000). Strategyproof assignment by hierarchical exchange. *Econometrica*, *68*(6), 1403–1433.

Pathak, P. A., & Sönmez, T. (2013). School admissions reform in Chicago and England: Comparing mechanisms by their vulnerability to manipulation. *American Economic Review*, *103*(1), 80–106.

Satterthwaite, M. A., & Sonnenschein, H. (1981). Strategy-proof allocation mechanisms at differentiable points. *Review of Economic Studies*, *48*(4), 587–597.

Shapley, L., & Scarf, H. (1974). On cores and indivisibility. *Journal of Mathematical Economics*, *1*(1), 23–37.

Szobolits, A., & Sunjic, M. H. (2007). Many refugees misunderstand Schengen expansion. *UN High Commissioner for Refugees (UNHCR)*.

Journal of Mechanism and Institution Design
ISSN: 2399-844X (Print), 2399-8458 (Online)
DOI:10.22574/jmid.2022.12.004

CHARACTERIZATION OF INCENTIVE COMPATIBLE SINGLE-PARAMETER MECHANISMS REVISITED

Krzysztof R. Apt

CWI, Amsterdam, The Netherlands
MIMUW, University of Warsaw, Poland
apt@cwi.nl

Jan Heering

CWI, Amsterdam, The Netherlands (retired)
jan.heering1@gmail.com

ABSTRACT

We reexamine the characterization of incentive compatible single-parameter mechanisms introduced in Archer & Tardos (2001). We argue that the claimed uniqueness result, called 'Myerson's Lemma' was not well established. We provide an elementary proof of uniqueness that unifies the presentation for two classes of allocation functions used in the literature and show that the general case is a consequence of a little known result from the theory of real functions. We also clarify that our proof of uniqueness is more elementary than the previous one. Finally, by generalizing our characterization result to more dimensions, we provide alternative proofs of revenue equivalence results for multiunit auctions and combinatorial auctions.

Keywords: Incentive compatibility, single-parameter mechanisms, Myerson's lemma, auctions, revenue equivalence.

JEL Classification Numbers: D44, D82.

We thank Guido Schäfer for suggesting to analyze the characterization result discussed in this paper and Marcin Dziubiński for helpful comments. We are grateful to one of the referees for recommending to discuss the uniqueness proofs given in Krishna (2009) and Börgers (2015).

1. INTRODUCTION

We are concerned here with a characterization result of a specific class of incentive compatible direct selling mechanisms. For the sake of this article such a *mechanism* consists of an *allocation rule* that assigns some good or goods to the participants, called agents, and a *payment rule* that determines how much each agent needs to pay. Assuming that each agent has a *private valuation* of the good or goods, these decisions are taken in response to a vector of *bids* made by the agents. These bids may differ from agents' true valuations. Recall that a mechanism is (*dominant strategy*) *incentive compatible* (alternatively called *truthful*), if no agent is better off when providing false information regardless of reports of the other agents or, more precisely, when submitting a bid different from his/her valuation regardless of reports of the other agents.

Given a class of mechanisms one of the main problems is to characterize their incentive compatibility in terms of an appropriate payment rule. Several such results were established in the literature, starting with the one in Green & Laffont (1977) concerning Groves mechanisms, originally proposed in Groves (1973). One of the earliest characterization results was given in Myerson (1981), who considered single object auctions in an imperfect information setting. In Milgrom (2004) such characterizations are called 'Myerson's Lemma'. This terminology was adopted in Roughgarden (2016), Chapter 3 of which, titled 'Myerson's Lemma', is concerned with a characterization of incentive compatible single-parameter mechanisms, which were studied in Archer & Tardos (2001).

As we explain below, both in this article and in Roughgarden's book such a characterization result is actually not proved. Most (but not all) of the claims are rigorously established in Nisan (2007) in the context of randomized single-parameter mechanisms.

Given that this purported characterization of incentive compatible single-parameter mechanisms is frequently referred to in the literature (see e.g. Hartline & Karlin (2007) and Babaioff (2016)), we find it justified to review these claims. We will provide an elementary proof of the characterization result for two classes of allocation functions considered in Roughgarden (2016) and subsequently provide a proof of the original claim of Archer & Tardos (2001) by appealing to more advanced results from the theory of real functions. We conclude by comparing our proof to the one given in Krishna (2002) and

Börgers (2015).

2. PRELIMINARIES

We follow here the terminology of Roughgarden (2016) that is slightly different than the one originally used in Archer & Tardos (2001). In particular Roughgarden (2016) refers to a single-parameter mechanism and an allocation rule, while Archer & Tardos (2001) refer to a one-parameter mechanism and load.

Each *single-parameter mechanism* concerns sale of some 'stuff' to bidders and assumes

- a set of agents $\{1,\ldots,n\}$,
- for every agent i, a value $v_i \geq 0$ which specifies i's *private valuation* "per unit of stuff" that he or she acquires.

In the auction the agents simultaneously submit their *bids*, which are their reported valuations "per unit of stuff". The auctioneer receives the bids and determines how much 'stuff' each agent receives and against which price. So in contrast to the single-item auctions each agent i receives a possibly fractional amount $a_i \geq 0$ of an object (here 'a stuff') he or she is interested in.

An *allocation* is a vector $\mathbf{a} = (a_1,\ldots,a_n)$, where each $a_i \geq 0$ specifies the amount allocated to agent i. A *payment* is a vector $\mathbf{p} = (p_1,\ldots,p_n)$, where each $p_i \geq 0$ specifies the amount agent i has to pay.

Each single parameter mechanism consists of an *allocation rule*

$$\mathbf{a} : \mathbb{R}^n_{\geq 0} \to \mathbb{R}^n_{\geq 0}$$

and a *payment rule*

$$\mathbf{p} : \mathbb{R}^n_{\geq 0} \to \mathbb{R}^n_{\geq 0}.$$

Given a vector of bids $\mathbf{b} = (b_1,\ldots,b_n)$ such a mechanism selects an allocation $\mathbf{a}(\mathbf{b}) = (a_1(\mathbf{b}),\ldots,a_n(\mathbf{b}))$ and a vector of payments $\mathbf{p}(\mathbf{b}) = (p_1(\mathbf{b}),\ldots,p_n(\mathbf{b}))$.

We assume that the *utility* of agent i is defined by

$$u_i(\mathbf{b}) = v_i a_i(\mathbf{b}) - p_i(\mathbf{b}).$$

We then say that a single-parameter mechanism is *incentive compatible* if for each agent i truthful bidding, i.e., bidding v_i, yields the best outcome

regardless of bids of the other agents. More formally, it means that for all agents i

$$u_i(v_i, \mathbf{b}_{-i}) \geq u_i(b_i, \mathbf{b}_{-i}),$$

for all bids b_i of agent i and all vectors of bids $\mathbf{b}_{-i}$ of other agents, or equivalently—ignoring the parameters $\mathbf{b}_{-i}$—that for all $y \geq 0$

$$v_i a_i(v_i) - p_i(v_i) \geq v_i a_i(y) - p_i(y).$$

3. A CHARACTERIZATION RESULT

We say that a function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is *monotonically non-decreasing*, in short *monotone*, if

$$0 \leq x \leq y \to f(x) \leq f(y).$$

We say that an allocation rule $\mathbf{a}$ is *monotone* if for every agent i and every vector of bids $\mathbf{b}_{-i}$ of other agents the function $a_i(\cdot, \mathbf{b}_{-i})$ is monotone.

The following result is stated in Archer & Tardos (2001), Nisan (2007), and Roughgarden (2016). In Nisan (2007) it is formulated as a result about randomized single-parameter mechanisms but the proofs are the same for the deterministic mechanisms considered here.

Theorem 1.

(i) If a mechanism $(\mathbf{a}, \mathbf{p})$ *is incentive compatible then the allocation rule* $\mathbf{a}$ *is monotone.*

(ii) If the allocation rule $\mathbf{a}$ *is monotone then for some payment rule* $\mathbf{p}$ *the mechanism* $(\mathbf{a}, \mathbf{p})$ *is incentive compatible.*

(iii) If the allocation rule $\mathbf{a}$ *is monotone, then all payment rules* $\mathbf{p}$ *for which the mechanism* $(\mathbf{a}, \mathbf{p})$ *is incentive compatible differ by a constant.*

There are some technically irrelevant differences between these three references. In Archer & Tardos (2001), instead of allocations, loads are considered, with the consequence that the loads are monotonically non-increasing, though the authors also state that the results equally apply to the set up that uses allocations. Following Nisan (2007) and Roughgarden (2016), we will use allocations. It leads to an analysis of monotonically non-decreasing functions. Further, in the last two references it is assumed that the payment rule

yields 0 payment when bids are equal to 0, which makes it possible to drop in (iii) the qualification 'up to a constant'. To make the discussion applicable to arbitrary payment rules we do not adopt this assumption.

Item (i) is established in Archer & Tardos (2001) by appealing to the first derivative, so under some assumptions about the load function. However, a short argument given in Nisan (2007) and reproduced in Roughgarden (2016) shows that no assumptions are needed.

In turn, item (ii) is proved in Archer & Tardos (2001) 'by picture'. A rigorous proof is given in Nisan (2007), while in Roughgarden (2016) only a 'proof by picture' is provided for piecewise constant allocation rule and it is mentioned that "the same argument works more generally for monotone allocation rules that are not piecewise constant".

Finally, in Archer & Tardos (2001) item (iii) is claimed for arbitrary monotone loads and allocation rules. But in the paper only a short proof sketch is given that ends with a claim that "To prove that all truthful payment schemes take form (2), even when ω_i [the load rule] is not smooth, we follow essentially the same reasoning as in the [earlier given] calculus derivation." However, this derivation refers to load rules that are assumed to be smooth (actually only twice differentiable, so that integration by parts can be applied), while the characterization result is claimed for all monotone allocation functions.

In Nisan (2007) item (iii) is established by reducing in the last step the expression $\int_0^x zf'(z)dz$ to $xf(x) - \int_0^x f(z)dz$. We quote (adjusting the notation): "[...] we have that $p(x) = \int_0^x zf'(z)dz$, and integrating by parts completes the proof. (This seems to require the differentiability of f, but as f is monotone this holds almost everywhere, which suffices since we immediately integrate.)" (Recall that a property holds *almost everywhere* if it holds everywhere except at a set of measure 0, i.e., a set that can be covered by a countable union of intervals the total length of which is arbitrarily small.) A minor point is that the initial part of the proof is incomplete as it only deals with the right-hand derivative instead of the derivative.

Finally, in Roughgarden (2016) about item (iii) it is only stated without proof "We reiterate that these payments formulas [for the above two classes of allocation functions] give *the only possible* payment rule that has a chance of extending the given allocation rule $\mathbf{x}$ into a DSIC [i.e., incentive compatible] mechanism." Also here the formula (in the adjusted notation) $p(x) = \int_0^x zf'(z)dz$ is derived by discussing only the right-hand derivative.

In our view these arguments are incomplete as they do not take into account some restrictions that need to be imposed on the use of integrals and application of integration by parts. Note that, except in the final discussion, Riemann integration is assumed throughout.

Remark 2. To start with, integration by parts can fail for simple monotone functions, for example those considered in Roughgarden (2016). Indeed, let for $q > 0$

$$H_q(x) := \begin{cases} 0 & \text{if } 0 \le x \le q \\ 1 & \text{if } x > q \end{cases}$$

be an elementary step function with a single step at $x = q$.

Take now $f = H_q$ with $q = 1/2$. Then $f' = 0$ for $x \neq 1/2$ and f' is undefined for $x = 1/2$. Consequently (defining $f'(1/2)$ arbitrarily)

$$\int_0^1 zf'(z)dz = 0 \neq 1/2 = xf(x)\big|_0^1 - \int_0^1 f(z)dz. \tag{1}$$

Further, integration by parts can fail even if we insist on continuity. Indeed, take for f the Cantor function, see, e.g., (Tao, 2011, pages 170-171). It is monotone, continuous and almost everywhere differentiable on $[0,1]$, with $f(0) = 0$, $f(1) = 1$ and f' equal to 0 whenever defined. Additionally, (1) holds for f, as well.

Finally, there exists a monotone and everywhere differentiable function f for which the above integral $\int_0^1 zf'(z)dz$ does not exist. Indeed, as observed in Goffman (1977), there exists a monotone and everywhere differentiable function $f : [0,1] \to \mathbb{R}_{\ge 0}$ such that the integral $\int_0^1 f'(z)dz$ does not exist. By a result of Lebesgue (see, e.g., Bressoud (2008)) a bounded function defined on a bounded and closed interval is Riemann integrable iff it is continuous almost everywhere. But f' is continuous almost everywhere on $[0,1]$ iff the function $g(x) := xf'(x)$ is, so the claim follows. □

These points of concern motivate our subsequent considerations. To keep the paper self-contained we reprove items (i) and (ii), given that the proofs are very short.

4. AN ANALYSIS

Our analysis can be carried out without any reference to mechanisms by reasoning about functions on reals. We first rewrite the incentive compatibility

condition as

$$p_i(y) - p_i(v_i) \geq v_i(a_i(y) - a_i(v_i)),$$

which from now on we analyze as the following condition on two functions $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$:

$$\forall x, y : g(y) - g(x) \geq x(f(y) - f(x)). \tag{2}$$

We are interested in solutions in g given f. We begin with the following obvious observation.

Note 3. *The inequality* (2) *is equivalent to*

$$\forall x, y : y(f(y) - f(x)) \geq g(y) - g(x) \geq x(f(y) - f(x)). \tag{3}$$

Proof. By interchanging in (2) x and y we get the additional inequality $y(f(y) - f(x)) \geq g(y) - g(x)$. □

Corollary 4 (Nisan (2007); Roughgarden (2016))**.** *Suppose* (2) *holds. Then the function f is monotone.*

Proof. Assume $0 \leq x < y$. By Note 3 (3) holds. The inequalities in (3) imply $(y-x)(f(y) - f(x)) \geq 0$, so $f(x) \leq f(y)$. □

This establishes item (i) of Theorem 1. To investigate items (ii) and (iii) we study existence and uniqueness of solutions of (2) in g. The following result establishes item (ii). The proof is from Nisan (2007).

Lemma 5. *Suppose f is monotone. Then* (2) *holds for*

$$g(x) = C + xf(x) - \int_0^x f(z)dz, \tag{4}$$

where C is some constant.

Because f is monotone g is well defined (see, e.g., Rudin (1976)).

Proof. By plugging the definition of g in (2) we get after some simplifications

$$\int_0^x f(z)dz - \int_0^y f(z)dz \geq (x-y)f(y), \tag{5}$$

which needs to be proved. Two cases arise.

Case 1. $x \geq y$.

Then

$$\int_0^x f(z)dz - \int_0^y f(z)dz = \int_y^x f(z)dz \geq (x-y)f(y),$$

where the last step follows by bounding the integral from below, since by the monotonicity of f, we have $f(y) \leq f(z)$ for $z \in [y,x]$.

Case 2. $y > x$.

Then

$$\int_0^x f(z)dz - \int_0^y f(z)dz = -\int_x^y f(z)dz \geq (x-y)f(y),$$

where the last step follows by bounding the integral from above, since by the monotonicity of f, we have $f(z) \leq f(y)$ for $z \in [x,y]$.

So (5) holds, which concludes the proof. □

To deal with uniqueness let us first consider the case for which the argument given in Nisan (2007) can be justified.

Lemma 6. *Suppose f is everywhere differentiable. Then any two solutions g of (2) differ by a constant.*

Proof. Suppose that (2) holds. By Note 3 (3) holds. Given an arbitrary $x \geq 0$ we first use it with $y = x+h$, where $h > 0$. Dividing by h we then obtain

$$\frac{(x+h)(f(x+h)-f(x))}{h} \geq \frac{g(x+h)-g(x)}{h} \geq \frac{x(f(x+h)-f(x))}{h}.$$

By the assumption about f

$$\lim_{h\to 0^+} \frac{(x+h)(f(x+h)-f(x))}{h} = \lim_{h\to 0^+} \frac{x(f(x+h)-f(x))}{h} = xf'(x),$$

so

$$\lim_{h\to 0^+} \frac{g(x+h)-g(x)}{h} = xf'(x). \tag{6}$$

Next, we use (3) with $x = y+h$, where $h < 0$. Dividing by h we then obtain

$$\frac{y(f(y)-f(y+h))}{h} \leq \frac{g(y)-g(y+h)}{h} \leq \frac{(y+h)(f(y)-f(y+h))}{h},$$

so replacing y by x and multiplying by -1 we get

$$\frac{x(f(x+h)-f(x))}{h} \geq \frac{g(x+h)-g(x)}{h} \geq \frac{(x+h)(f(x+h)-f(x))}{h}.$$

By the assumption about f and x

$$\lim_{h\to 0^-} \frac{(x+h)(f(x+h)-f(x))}{h} = \lim_{h\to 0^-} \frac{x(f(x+h)-f(x))}{h} = xf'(x),$$

so

$$\lim_{h\to 0^-} \frac{g(x+h)-g(x)}{h} = xf'(x). \tag{7}$$

We conclude from (6) and (7) that $g'(x)$ exists and

$$g'(x) = xf'(x). \tag{8}$$

Hence all solutions g to (2) have the same derivative and consequently differ by a constant. □

Remark 7. The above proof coincides with the one given in Nisan (2007), except on two points. First, only (6) is established there. This allows one only to conclude that the right derivative of g in x exists; to establish that $g'(x)$ exists also (7) is needed. More importantly, Nisan argued that all solutions g to (2) are of the form (4) given in Lemma 5. Under the assumption that f is everywhere differentiable this additional claim is a direct consequence of Lemmas 5 and 6.

Nisan's argument for this point involved integration and integration by parts. To justify it we need to assume that f' is continuous. Then by (8) also g' is continuous, which allows us to use the Fundamental Theorem of Calculus. It yields that for some constant C

$$g(x) = C + \int_0^x g'(z)dz.$$

Further, integration by parts of $\int_0^x zf'(z)dz$ is then also justified since f is everywhere differentiable and f' is integrable (see, e.g., Rudin (1976)). Then $\int_0^x zf'(z)dz$ exists and by integration by parts

$$\int_0^x zf'(z)dz = xf(x) - \int_0^x f(z)dz,$$

so (8) and the last two equalities imply that g is indeed of the form (4) given in Lemma 5. □

In the remainder of this section we do not use integration or the existence of solutions in the form (4), but proceed directly from (2). This allows us to sidestep the associated complications and show that the requirement of f being everywhere differentiable of Lemma 6 can be substantially weakened and, appealing to strong results from the theory of real functions, can even be removed altogether.

For $x = 0$ continuity (differentiability) means right continuity (differentiability), which we will not mention or treat separately.

We first need an auxiliary result.

Lemma 8. *Let g_1 and g_2 be two solutions of (2) and let $G = g_1 - g_2$.*

(i) G is continuous.

(ii) If f is continuous at x, then G is differentiable at x and $G'(x) = 0$.

Proof. (i) By Note 3 (3) holds for g_1 and g_2. By using it with $y = x+h$ for g_1 and for g_2 we obtain

$$0 \leq |G(x+h) - G(x)| \leq h(f(x+h) - f(x)). \tag{9}$$

We have $h(f(x+h) - f(x)) \leq hf(x+h)$ for $h > 0$ and $\leq -hf(x)$ for $h < 0$. But by Corollary 4 f is monotone, hence $\lim_{h\to 0} |G(x+h) - G(x)| = 0$, which establishes the claim.

(ii) Take some $x \geq 0$. By (9) for $h \neq 0$

$$0 \leq \left| \frac{G(x+h) - G(x)}{h} \right| \leq |f(x+h) - f(x)|,$$

which implies the claim. □

Note that the continuity of G holds for any f.

The following result with an elementary proof covers in a unified way item (iii) of Theorem 1 for two classes of allocation functions considered in Roughgarden (2016), piecewise constant and differentiable ones.

A function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is called *piecewise continuous* if it has at most a finite number of discontinuities in every bounded interval. Thus discontinuities can occur only at isolated points separated by open intervals of continuity. Piecewise constant and step functions are special cases. This definition is a straightforward generalization to $\mathbb{R}_{\geq 0}$ of the usual one for functions with a bounded domain.

Theorem 9. *Suppose f is piecewise continuous. Then any two solutions g of (2) differ by a constant.*

Proof. Let g_1 and g_2 be two solutions of (2) and let $G = g_1 - g_2$.

Let f be piecewise continuous with discontinuities $q_1 < q_2 < \ldots$ and consider the intervals $I_0 = [0, q_1)$ ($\emptyset$ if $q_1 = 0$), $I_i = (q_i, q_{i+1})$ ($i \geq 1$), with $I_N = (q_N, \infty)$ if f has a finite number $N > 0$ of discontinuities and $I_0 = [0, \infty) = \mathbb{R}_{\geq 0}$ if $N = 0$.

If f has an infinite number of discontinuities, $\lim_{i \to \infty} q_i = \infty$ since there can be only finitely many of them in any bounded interval. Hence, the I_i and q_i together cover the whole of $\mathbb{R}_{\geq 0}$.

f is continuous on each I_i, so by Lemma 8(*ii*) G is constant on I_i, say $G = C_i$ on I_i. Since G is continuous everywhere by Lemma 8(*i*), $C_0 = G(q_1) = C_1 = G(q_2) = \ldots$, so for some constant C, $G = g_1 - g_2 = C$. □

Theorem 9 can be generalized to a wider class of functions whose discontinuity sets may have limit points (accumulation points), at least to some degree. We give a simple example of a monotone function f, for which Theorem 9 does not apply but the stronger result presented below does.

Let

$$f(x) = \sum_{n=1}^{\infty} 2^{-n} H_{1-2^{-n}}(x),$$

where H_q is defined in Remark 2. It is not piecewise continuous, but has an infinite set of discontinuities $\{1/2, 3/4, \ldots\}$ with a single limit point 1. Note that f happens to be continuous at $x = 1$, but this might also have been otherwise.

Functions like f and more complicated ones having discontinuity sets with limit points of limit points, etc. can, to some extent, be dealt with by adapting the proof of Theorem 9 and appealing to the well-known Bolzano-Weierstrass theorem (BW for short, see, e.g., Bressoud (2008)).

Given a set $S \subseteq \mathbb{R}_{\geq 0}$, we denote by $S^{(1)}$ the set of its limit points (which need not be in S) and define $S^{(n+1)} = (S^{(n)})^{(1)}$ for $n \geq 1$. A set S is called *first species* of type $n-1$ if $S^{(n)} = \emptyset$ and $S^{(m)} \neq \emptyset$ for $m < n$ (Bressoud, 2008). Such a set has limit points, limit points of limit points, etc., up to level $n-1$. A first species set of type 0 has no limit points.

Theorem 10. *Suppose the discontinuity set of f is first species of type $n \geq 0$. Then any two solutions g of (2) differ by a constant.*

Proof. Let g_1 and g_2 be two solutions of (2) and let $G = g_1 - g_2$. Let the discontinuity set of f be S and use induction with respect to the type of S.

If $n = 0$, $S^{(1)} = \emptyset$, so f can have only finitely many discontinuities in every bounded interval. Otherwise, there would be a limit point in some bounded and closed interval by BW. Hence, f is piecewise continuous and the case $n = 0$ corresponds to Theorem 9.

Assume the theorem holds for all $n \leq k$ for some $k > 0$ and consider S of type $k+1$. The elements of $S^{(k+1)}$ are the limit points of level $k+1$. Recall that these need not be elements of S. Since $S^{(k+2)} = \emptyset$, $S^{(k+1)}$ does not have limit points, so there can be only finitely many elements of $S^{(k+1)}$ in every bounded interval by BW as before.

Now let the elements of $S^{(k+1)}$ be $q_1 < q_2 < \ldots$ and consider the intervals I_i ($i \geq 0$) as in the proof of Theorem 9. Together with the q_i, they cover the whole of $\mathbb{R}_{\geq 0}$ as in the previous proof.

However, the $S \cap I_i$ may still have q_i and/or q_{i+1} as limit points, which means the $S \cap I_i$ need not be of type $\leq k$ but can still be of type $k+1$, so the induction hypothesis cannot be applied to the I_i. Therefore, for fixed i and sufficiently small $\delta > 0$ consider a non-empty bounded and closed subinterval $J_i = J_i(\delta) = [q_i + \delta, q_{i+1} - \delta]$ of I_i (or, if the number of limit points is a finite number N, $J_N = J_N(\delta) = [q_N + \delta, \infty)$).

Now $S \cap J_i$ can no longer have q_i and/or q_{i+1} as limit points. Hence, it is of type $\leq k$ and the induction hypothesis applies to J_i, so G is constant on J_i, say $G = C_i$ on J_i. Since G is continuous everywhere by Lemma 8(i) and $\lim_{\delta \to 0} J_i = [q_i, q_{i+1}]$ (or, if the number of limits points is N, $\lim_{\delta \to 0} J_N = [q_N, \infty)$), we get $G(q_i) = C_i = G(q_{i+1})$. The final step of the proof is the same as in the proof of Theorem 9. □

Unfortunately, the above result does not cover all monotone functions $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$. Indeed, a monotone function may be discontinuous on the set $\mathbb{Q}_{\geq 0}$ of non-negative rational numbers, see, e.g., Rudin (1976), and $\mathbb{Q}_{\geq 0}$ is not first species, since $\mathbb{Q}_{\geq 0}^{(1)} = \mathbb{R}_{\geq 0}$ and $\mathbb{R}_{\geq 0}^{(1)} = \mathbb{R}_{\geq 0}$.

This limitation can be circumvented by appealing to a strong result of Goldowski and Tonelli. Recall first that a function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is differentiable *nearly everywhere* if it is differentiable except at a countable number of points. Note that *nearly everywhere* implies *almost everywhere*. We need

Theorem 11 (Goldowski (1928); Tonelli (1930-31); Saks (1937))**.** *Let* $G : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ *be a function such that*

- *G is continuous,*
- *G is differentiable nearly everywhere,*
- *$G' \geq 0$ almost everywhere.*

Then G is monotone.

This leads directly from Lemma 8 to the desired conclusion.

Theorem 12. *Any two solutions g of (2) differ by a constant.*

Proof. Let g_1 and g_2 be two solutions of (2) and let $G = g_1 - g_2$. By Lemma 8(i) G is continuous.

A monotone function is continuous nearly everywhere (see, e.g., Rudin (1976)). So by Lemma 8(ii) G is differentiable nearly everywhere and $G' = 0$ nearly everywhere. Hence by Theorem 11 both G and $-G$ are monotone, i.e., G is constant. □

The above theorem justifies item (iii) of Theorem 1. The following result summarizes the results of this section.

Theorem 13. *Inequality (2) holds iff f is monotone and for some constant C*

$$g(x) = C + xf(x) - \int_0^x f(z)\,dz.$$

Proof. By Corollary 4, Lemma 5, and Theorem 12. □

5. DISCUSSION

Results closely corresponding to our uniqueness result (Theorem 12) were also presented in Krishna (2002) (and its second edition Krishna (2009)) and Börgers (2015). The customary name of these results is Revenue Equivalence. Krishna considers in Chapter 5 a setup with a seller that has one indivisible object to sell and n potential buyers, while Börgers considers in Chapter 2 a setup in which there is just one potential buyer. In Krishna (2009) the equivalent of our function f is defined as an integral representing the probability that a buyer gets the object, while in Börgers (2015) f corresponds to the probability of selling the object to the buyer. However, a close inspection of the

proofs of these Revenue Equivalence results reveals that they do not depend on the actual form of f.

Further, ignoring the differences in the setup, the corresponding proofs in both books are from the mathematical point of view essentially the same. As the arguments in the latter one are more detailed, we discuss them here, but using our notation.

The proof of the corresponding result (Proposition 2.2) in Börgers (2015) is not based on the equivalent of our function f but instead deals, in Lemma 2.2, with the function u (representing utility) defined by

$$u(x) := xf(x) - g(x),$$

and states that for all x for which u is differentiable,

$$u'(x) = f(x).$$

Lemma 2.2 also establishes that the function u is monotone and convex. Then in Lemma 2.3 it is shown that

$$u(x) = u(0) + \int_0^x f(z)dz,$$

which is equivalent to (4) by taking $C = u(0)$, so the uniqueness result (Lemma 2.4 (Revenue Equivalence)) corresponding to our Theorem 12, follows.

Lemma 2.3 is a direct consequence of two results from Royden & Fitzpatrick (2010), namely, that convexity implies absolute continuity (a notion we leave undefined here) and that every absolutely continuous function is equal to the integral of its derivative.

Note, however, that the latter result (Theorem 10 of Royden & Fitzpatrick (2010)) is the Fundamental Theorem of Calculus (FTC) for the Lebesgue integral, a fact not mentioned in Krishna (2002) and in Börgers (2015) deducible only indirectly from footnote 2 in Chapter 2. So the proofs of the uniqueness result (the Revenue Equivalence) presented in Krishna (2002) and Börgers (2015) *crucially* rely on the Lebesgue theory of integration. In contrast, our proof is much more elementary: it does not rely on *any* form of integration and appeals only to the notion of derivative. Only the existence result (Lemma 5) relies on the Riemann integral. Having said this, apart from its complications, use of the Lebesgue integral yields a very efficient proof.

Both Krishna and Börgers establish appropriate Revenue Equivalence results for other mechanisms. In particular, Börgers considers in Chapters 3 and

4 of Börgers (2015) Bayesian mechanisms and dominant mechanisms, each time for n buyers. In both cases he establishes the corresponding Revenue Equivalence result (Lemma 3.4 and Proposition 4.2) by explaining that the reasoning provi/ded in Chapter 2 can be repeated.

Given that in both setups the crucial inequalities are the counterparts of (2) (considered separately for each buyer), it follows that both results can be alternatively proved using our Theorem 12. We conclude that our more elementary approach can be applied to other mechanisms than the single-parameter mechanism considered in Section 2.

Finally, we show that a generalization of Theorem 13 allows one to provide alternative, more elementary proofs of Revenue Equivalence for two other types of auctions, considered in Krishna (2009) in Chapters 14 and 16. These auctions are concerned with multiple objects, leading to functions f and g having to deal with vectors.

We use the following notation. For functions $f : \mathbb{R}^n_{\geq 0} \to \mathbb{R}^n_{\geq 0}$ and $g : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$ and $\mathbf{x} \in \mathbb{R}^n_{\geq 0}$ we introduce the functions $f_{\mathbf{x}} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ and $g_{\mathbf{x}} : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ defined by

$$f_{\mathbf{x}}(t) = f(t\mathbf{x}) \cdot \mathbf{x},$$

where $\cdot$ is the inner product (dot product), and

$$g_{\mathbf{x}}(t) = g(t\mathbf{x}).$$

We then say that $f : \mathbb{R}^n_{\geq 0} \to \mathbb{R}^n_{\geq 0}$ is *monotone* if for each $\mathbf{x} \in \mathbb{R}^n_{\geq 0}$ the function $f_{\mathbf{x}}$ is monotone.

We now establish the following result which generalizes Theorem 13 to dimension $n > 1$. (Note that using the substitution $u(z) = zx$ we have $\int_0^x f(u)\,du = \int_0^1 f(zx)x\,dz$.)

Theorem 14. For two functions $f : \mathbb{R}^n_{\geq 0} \to \mathbb{R}^n_{\geq 0}$ and $g : \mathbb{R}^n_{\geq 0} \to \mathbb{R}_{\geq 0}$ the inequality

$$\forall \mathbf{x}, \mathbf{y} : g(\mathbf{y}) - g(\mathbf{x}) \geq (f(\mathbf{y}) - f(\mathbf{x})) \cdot \mathbf{x} \tag{10}$$

holds iff f is monotone and for some constant C

$$g(\mathbf{x}) = C + f(\mathbf{x}) \cdot \mathbf{x} - \int_0^1 f(z\mathbf{x}) \cdot \mathbf{x}\,dz. \tag{11}$$

Proof. First note that (10) holds iff

$$\forall \mathbf{x}, x, y : g_{\mathbf{x}}(y) - g_{\mathbf{x}}(x) \geq x(f_{\mathbf{x}}(y) - f_{\mathbf{x}}(x)), \tag{12}$$

since for each $\mathbf{x} \in \mathbb{R}^n_{\geq 0}$

$$\forall x, y : g_{\mathbf{x}}(y) - g_{\mathbf{x}}(x) = g(y\mathbf{x}) - g(x\mathbf{x})$$

and by linearity of the inner product

$$\forall x, y : x(f_{\mathbf{x}}(y) - f_{\mathbf{x}}(x)) = (f(y\mathbf{x}) - f(x\mathbf{x})) \cdot x\mathbf{x}.$$

By Theorem 9 (12) holds iff for all $\mathbf{x}$ the function $f_{\mathbf{x}}$ is monotone and for some constant $C_{\mathbf{x}}$

$$g_{\mathbf{x}}(x) = C_{\mathbf{x}} + x f_{\mathbf{x}}(x) - \int_0^x f_{\mathbf{x}}(z)\, dz. \tag{13}$$

But $C_{\mathbf{x}} = g_{\mathbf{x}}(0) = g(\mathbf{0})$, so the constant $C_{\mathbf{x}}$ does not depend on $\mathbf{x}$. Further, $g(\mathbf{x}) = g_{\mathbf{x}}(1)$, and $1 f_{\mathbf{x}}(1) = f(\mathbf{x}) \cdot \mathbf{x}$, so by putting $x = 1$ we see that(13) implies (11).

But also (11) implies (13), which can be seen by using (11) with $x\mathbf{x}$ instead of $\mathbf{x}$.

□

Let us return now to Krishna (2009). In Chapter 14 he studies multiunit auctions in which multiple identical objects are available. The relevant inequality (14.1) on page 204, capturing the expected payment in an equilibrium for a player, corresponds to (10). Theorem 14 then provides an alternative proof of his Proposition 14.1 stating that

> "The equilibrium payoff (and payment) functions of any bidder in any two multiunit auctions that have the same allocation rule differ at most by an additive constant."

Krishna's proof relies (implicitly) on the Lebesgue integral. We adopted from his proof the idea of reasoning about the functions $f_{\mathbf{x}}$ and $g_{\mathbf{x}}$. In Chapter 16 of his book he studies auctions in which one can bid for a set of nonidentical objects. They are called in the computer science literature combinatorial auctions. Krishna explains that "The proof is *identical* to that of Proposition 14.1." (italic used by the author). Consequently, our approach also yields an alternative proof of Revenue Equivalence for combinatorial auctions.

References

Archer, A., & Tardos, É. (2001). Truthful mechanisms for one-parameter agents. In *IEEE Symposium on Foundations of Computer Science (FOCS)* (pp. 482–491). IEEE Computer Society.

Babaioff, M. (2016). Truthful mechanisms for one-parameter agents. In *Encyclopedia of Algorithms* (pp. 2267–2271). Springer.

Börgers, T. (2015). *An Introduction to the Theory of Mechanism Design*. Oxford University Press.

Bressoud, D. (2008). *A Radical Approach to Lebesgue's Theory of Integration*. Cambridge University Press.

Goffman, C. (1977). Bounded derivative which is not Riemann integrable. *The American Mathematical Monthly*, *84*(3), 205-206.

Goldowski, G. (1928). Note sur les dérivées exactes. *Rec. Math. Soc. Math. Moscou (Mat. Sbornik)*, *35*, 35-36.

Green, J. R., & Laffont, J. J. (1977). Characterization of satisfactory mechanism for the revelation of preferences for public goods. *Econometrica*, *45*, 427-438.

Groves, T. (1973). Incentives in teams. *Econometrica*, *41*, 617–631.

Hartline, J. D., & Karlin, A. R. (2007). Profit maximization in mechanism design. In N. Nisan, T. Roughgarden, É. Tardos, & V. J. Vazirani (Eds.), *Algorithmic Game Theory* (p. 331-361). Cambridge University Press.

Krishna, V. (2002). *Auction Theory*. Academic Press.

Krishna, V. (2009). *Auction Theory* (second ed.). Academic Press.

Milgrom, P. (2004). *Putting Auction Theory to Work*. Cambridge University Press.

Myerson, R. (1981). Optimal auction design. *Mathematics of Operations Research*, *6*(1), 58-73.

Nisan, N. (2007). Introduction to mechanism design (for computer scientists). In N. Nisan, T. Roughgarden, É. Tardos, & V. J. Vazirani (Eds.), *Algorithmic Game Theory* (p. 209-241). Cambridge University Press.

Roughgarden, T. (2016). *Twenty Lectures on Algorithmic Game Theory*. Cambridge University Press.

Royden, H., & Fitzpatrick, P. (2010). *Real Analysis* (fourth ed.). Pearson.

Rudin, W. (1976). *Principles of Mathematical Analysis* (third ed.). McGraw-Hill.

Saks, S. (1937). *Theory of the Integral* (second revised ed.). New York: Hafner Publishing Company. (Reprinted by Hassell Street Press in 2021.)

Tao, T. (2011). *An Introduction to Measure Theory*. American Mathematical Society.

Tonelli, L. (1930-31). Sulle derivate esatte. *Mem. Istit. Bologna*, *8*, 13-15.

Journal of Mechanism and Institution Design
ISSN: 2399-844X (Print), 2399-8458 (Online)
DOI:10.22574/jmid.2022.12.005

A REGULATORY ARBITRAGE GAME: OFF-BALANCE-SHEET LEVERAGE AND FINANCIAL FRAGILITY

Dimitris Voliotis

University of Piraeus, Greece
dvoliotis@unipi.gr

ABSTRACT

This study examines a simple banking system in a game-theoretic framework wherein banks act as self-interested agents to maximize leverage at the expense of overall financial stability. The resultant strategic inefficiency raises concerns about how banks manage the "financial stability" good, which is appropriated into a "tragedy of the commons." We conceptualize the inefficiency using the price of anarchy introduced by Koutsoupias & Papadimitriou (2009). We seek the optimal regulatory framework that minimizes the price of anarchy or the degree of financial fragility.

Keywords: Financial fragility, congestion games, price of anarchy, best response potential.

JEL Classification Numbers: G21, G28, D53, C72.

1. INTRODUCTION

SINCE the early 2000s, banks and financial institutions have constructed or employed new instruments to increase leverage without violating regulatory rules. These instruments might not be readily evident on balance sheets

This work has been partly supported by the University of Piraeus Research Center. I am grateful to the editor and two anonymous referees for their helpful comments and constructive suggestions.

yet still expose financial institutions to credit risk. Typically, they include letters of credit, guarantees, operating leases, CDO's, swaps, and OTC derivatives. These off-balance sheet instruments are considered when assessing banks' exposure to credit risk but are subject to different treatment because of their special character.

Such off-balance-sheet leverage could allow banks to transfer credit risk to investors and clear room for new investment opportunities. However, financial institutions, especially banks, employ them not for better risk-sharing but to avoid costly capital buffers and circumvent regulatory requirements. Financial institutions mask credit risk from regulators and increase their risk exposures by undertaking off-balance sheet and other securitizations. This "regulatory arbitrage" compounded the 2008 global financial crisis and its aftermath (Acharya & Richardson, 2009).

The literature on "regulatory arbitrage" is far from new. Pavel & Phillis (1987), and Baer & Pavel (1988) find that lower capital ratios or more demanding regulatory capital are associated with higher levels of off-balance-sheet activities. Jones (2000) discusses the techniques used to undertake regulatory arbitrage and the difficulties faced by regulatory authorities. Breuer (2002) addresses the problem of measurement of off-balance-sheet leverage. It discusses the interaction between risk and off-balance-sheet leverage and calculates a modified capital ratio that incorporates the enhanced leverage implicit in off-balance-sheet securities.

Regulators undertake to assure that financial institutions remain sound, especially banks as the backbone of the financial system. However, their primary tools for ensuring financial stability -cash reserve ratios and capital adequacy ratios- are macro-prudential. That is, they are suited to tackle systemic financial risks.

Systemic financial risk is the risk that an event will trigger a loss of economic value or confidence in a substantial portion of financial system (GTen, 2001). In fact, for a financial system of high concentration, the collapse of a single financial institution suffices to trigger a systemic event. Hence, financial regulatory authorities should be able to assess and manage financial risks to maintain the effectiveness of the financial mechanisms. Moreover, financial risks inhibit banks from diverting high-powered money to higher-return investments, and the latter make banks more resilient to economic downturns at the sacrifice of leverage. Those suggest that institution-level issues warrant attention as well.

We introduce a simple banking system in a game-theoretic framework wherein banks act as self-interested agents, maximizing profits at the expense of overall financial stability. As a result, a "tragedy of the commons" emerges in which banks' strategic behavior produces inefficient outcomes. To measure inefficiency, we use the concept of "price of anarchy" introduced by Koutsoupias & Papadimitriou (2009) and employed by Moulin (2007) and Juarez (2006). In a broader sense, the game we study is a congestion game, a class of games that admit an ordinal potential function (Monderer & Shapley, 1996) that assures an equilibrium outcome. We aim to measure the maximum inefficiency occurring at equilibrium and use that information to measure financial fragility. We elaborate on the bounds of strategic inefficiency, viz. Vetta (2002), and Roughgarden (2006, 2012). Calculating the upper bound of strategic inefficiency reveals how detrimental banks' opportunistic behavior can be.

A further extension of the model incorporates bankers who intend to shake up the financial system to pursue speculative profit. Such behavior is often overlooked in theoretical models; nevertheless, it is customary for speculators to increase market liquidity and short the market on the downside. The Byzantine Generals Problem is an appropriate framework for introducing such destabilizing behavior in the financial system, which originally appeared in distributed systems literature (Lamport et al., 2019). The story behind Byzantine generals is a metaphor for a connected network of agents that must reliably communicate a common plan of action. However, among the loyal agents, some traitors undermine the agreement. The question that emerges is how tolerant the network (i.e., the financial system) is of the perverse incentives of "traitors" (i.e., speculators).

The rest of the paper proceeds as follows. Section 2 presents our model and its measure of strategic inefficiency. In addition, it associates the boundedness properties of inefficiency with the literature on generic cost-minimization games. Section 3 extends the model to bankers that benefit by destabilizing the financial system. Section 4 concludes.

2. MODEL

Suppose a one-shot game involving $I = \{1, \ldots, n, n+1\}$ players. The first $n \geq 2$ players are banks and the $n+1$ is a pseudo-player that stands for the financial regulatory authority (FRA). Each bank has a simple balance sheet

on which liabilities are deposits (D) and bank's capital. The deposit rate is zero ($r^D = 0$) for convenience. Assets are cash balances and a single asset (A) that generates a positive return ($r^m > 0$). Regard r^m as return on assets, which without loss of generality, we assume is identical across banks. The FRA has a decisive role in the game. As a tool of regulatory policy, it adopts a *capital adequacy ratio* ψ (a percentage of assets to be held in cash) and a *reserve requirement ratio* θ (a percentage of deposits to be held in cash). Banks eventually incur a "regulatory tax" amounting to the opportunity cost of holding reserves and capital that could be invested for a positive return in asset A. Foregone profits for bank $i \in I \setminus \{n+1\}$ attributable to regulatory tax are estimated as

$$RT_i = \psi \cdot A_i \cdot r^m + \theta \cdot D_i \cdot r^m = r^m(\psi \cdot A_i + \theta \cdot D_i).$$

Absent regulation, the bank could invest both reserved deposits $(\theta \cdot D_i)$ and reserved capital $(\psi \cdot A_i)$ in asset A and enjoy with certainty a positive return r^m.

Cost function

We assume that incidents of financial distress occur horizontally during which all banks suffer a haircut of ω percent. Hence, the objective of bank i is to minimize total cost that includes the regulatory tax and ω. We assume two specifications of total cost. First, for bank i and a proper subset of banks $S \subseteq I \setminus \{n+1\}$ the cost function is given by

$$C_i = \alpha_i[r^m(\psi \cdot A_i + \theta \cdot D_i)] + (1-\alpha_i)\frac{\sum_{j=1}^{\#S}(1-\alpha_j)}{n}\omega \cdot A_i,$$

which for $\alpha_i \in (0,1)$ and substituting RT_i becomes

$$C_i = \alpha_i RT_i + (1-\alpha_i)^2\frac{\omega A_i}{n} + (1-\alpha_i)\frac{\sum_{j=1}^{\#S\setminus\{i\}}(1-\alpha_j)}{n}\omega \cdot A_i. \qquad (1)$$

Bank i decides to circumvent part $(1-\alpha_i)$ of the regulatory tax by committing off-balance-sheet activities and to remit the remainder α_i. Hence, α_i is the strategic variable of bank i and determines how much regulatory tax it circumvents. The first term on the right in Eq.(1) is the cost of regulatory tax. The second and third terms denote the expected loss from financial distress.

If $S = \{j \in I \setminus \{n+1\} | s.t.\ \alpha_j < 1\}$ is the subset of banks that circumvent some or all of their regulatory tax, the probability of financial distress will be $\frac{\sum_{j=1}^{\#S}(1-\alpha_j)}{n}$ and the haircut for individual bank i will be ωA_i. As the number of evading banks (i.e. $\#S \to n$) and the regulatory tax evasion ($\alpha_i \to 0$) increase the probability of financial distress tends toward 1.

A drawback of this specification is that the bank becomes immune to systemic risk when fully complying with regulation ($\alpha = 1$). We can relax this strong assumption by assuming that all banks bear the cost in case of financial distress. A different specification to accommodate these conditions is

$$\begin{aligned} C_i &= \alpha_i[r^m(\psi \cdot A_i + \theta \cdot D_i)] + \frac{\sum_{j=1}^{\#S}(1-\alpha_j)}{n}\omega \cdot A_i \\ &= \alpha_i[r^m(\psi \cdot A_i + \theta \cdot D_i)] + \frac{(1-\alpha_i)}{n}\omega \cdot A_i + \frac{\sum_{j\neq i}(1-\alpha_j)}{n}\omega \cdot A_i. \end{aligned} \tag{2}$$

Banks reduce the probability of financial distress if they fully comply with financial regulation, but they can be *contaminated* by a financial crunch and suffer its consequences. We call this cost function the *cost with contagion effect*.

Price of Financial Anarchy

From the FRA's perspective the *social optimum cost* is that all banks opt for $\alpha_i = 1$. Doing so makes the overall cost equal to $\bar{\mathbf{C}} = \sum_i C_i((\psi, \theta, \alpha = \mathbf{1})$. *Social optimum cost* is the overall regulatory tax,

$$SOC = \bar{\mathbf{C}} = \sum_i RT_i = RT.$$

At Nash equilibrium overall (social) cost is denoted $\mathbf{C}^* = \sum_i C_i((\psi, \theta), \alpha^*)$. Departures from social optimum can be computed using a coordination ratio, known in game-theoretic literature as the *price of anarchy* (Koutsoupias & Papadimitriou, 2009). For our model, we call it the *price of financial anarchy* (PFA). The ratio of the cost at Nash equilibrium over the social optimum cost ($C^*/\bar{C}$) can be a metric of banking system disobedience to FRA policies.

Definition 1. *PFA is defined as the maximum deviation from social optimum cost for the worst-case equilibrium in the set of equilibria. It is the ratio*

$$PFA = \max_{\alpha^* \in NE} \frac{\mathbf{C}^*}{\bar{\mathbf{C}}}. \tag{3}$$

We anticipate the PFA metric to take values above 1. The following lemma proves the result.

Lemma 1. *PFA* > 1*, when for all banks i*

$$\frac{\omega A_i}{RT_i} \geq \frac{n}{\sum_{\forall j}(1-\alpha_j)}.$$

Proof. Appendix A.

Similar to PFA is the price of financial stability, which measures the deviation from the best-case equilibrium. The two measures coincide in the case of a unique equilibrium.

The Regulatory Arbitrage Game

To exert its stabilizing role in the financial system, the FRA aims to minimize the objective $C_{n+1} = |PFA - 1|$ by choosing appropriate policy parameters (ψ, θ). The game's strategy profile of banks is $\alpha \in [0,1]^n$ given the policy mix of the FRA .

Definition 2. *The regulatory arbitrage game is defined by the tuple*

$$\Gamma = \{I, \{[0,1]\}_{i\in I}, \{C_i\}_{i\in I}, (\psi, \theta)\}.$$

Nash equilibrium

Nash equilibrium emerges when all banks simultaneously minimize their total costs.

Definition 3. *Given policy parameters* (ψ, θ)*, a Nash equilibrium is a strategy profile* α^* *such that for all banks i*

$$C_i((\psi,\theta),\alpha^*)) \leq C_i((\psi,\theta),(\alpha_i,\alpha^*_{-i})) \quad \textit{for all} \;\; \alpha_i.$$

Remark. Nash equilibrium always exists. Both strategy sets and cost functions are convex, so a minimum always exists.

The regulatory arbitrage game admits a *best-response potential*. Best-response potential games guarantee Nash equilibrium when each player's cost

function is non-linear. The game $\Gamma = \{I, [0,1]^n, \{C_i\}_{i\in I}, (\psi, \theta)\}$ admits *best-response potential* $P : [0,1]^n \mapsto R$ such that

$$\arg\min_{\alpha_i} C_i(\alpha) = \arg\min_{\alpha_i} P(\alpha)$$

Next, we provide the best response potential for our game.

Lemma 2. *The best-response potential function of the regulatory arbitrage game is*

$$P(\alpha) = \sum_i (1-a_i)^2 \frac{\omega A_i}{n}.$$

Proof. Appendix B.

Now we provide the existence of equilibrium.

Proposition 1. *The Nash equilibrium of the regulatory arbitrage game always exists.*

Proof. By Lemma 2 and Proposition 2.2 in Voorneveld (2000) the game has a Nash equilibrium. □

Bounds to inefficiency

The finiteness and the noncooperative character of the game make the equilibrium outcome Pareto inefficient. It is true that for a noncooperative game with a finite and self-interested set of players, there is always a non-equilibrium outcome that is superior in a Pareto sense (Dubey, 1986). Exemplified games are the prisoner's dilemma and the standard Cournot duopoly; nevertheless, the degree of inefficiency remains to be found. For this task, we provide further definitions. We denote $\alpha_{-i} \geq \alpha'_{-i}$ whenever component-wise $\alpha_j \geq \alpha'_j$ for all $j \neq i$. With slight abuse of notation, also let $\alpha_{-i} = \alpha_j$.

Definition 4. *We say that the cost function exhibits decreasing differences if for $a_i \geq a'_i$ and $a_j \geq a'_j$ it is*

$$C_i(\alpha_i, \alpha_j) - C_i(\alpha_i, \alpha'_j) \leq C_i(\alpha'_i, \alpha_j) - C_i(\alpha'_i, \alpha'_j), \quad \forall i \in I \setminus \{n+1\}. \tag{4}$$

A game with cost functions that exhibit decreasing differences is *submodular*. Next, we show that bankers' cost functions exhibit decreasing differences.

Lemma 3. *The cost function of banks in the regulatory arbitrage game exhibits decreasing differences i.e., for* $\alpha_i \geq \alpha_i'$ *and* $\alpha_j \geq \alpha_j'$ *it is*

$$C_i(\alpha_i, \alpha_j) - C_i(\alpha_i, \alpha_j') \leq C_i(\alpha_i', \alpha_j) - C_i(\alpha_i', \alpha_j'), \quad \forall \in I \setminus \{n+1\}.$$

Proof. Appendix C.

Lemma 3 assures that banks always have escalating incentives to circumvent regulatory tax, for the greater the circumvention, the greater is the cost-saving. Next, we prove that the decreasing differences are linear for the second specification of the cost function. Finally, the next lemma says that submodularity is maintained, albeit linearly, under contagion effects.

Lemma 4. *The cost function with contagion effect exhibits linear decreasing differences. That is,*

$$C_i(\alpha_i, \alpha_j) - C_i(\alpha_i, \alpha_j') = C_i(\alpha_i', \alpha_j) - C_i(\alpha_i', \alpha_j'), \quad \forall i.$$

Proof. Appendix D.

Also, important in this analysis is individual bank's ability to affect social cost. Suppose the extreme wherein all banks except bank i comply fully with regulatory policy. That is (ψ, θ), ie $C_i((\psi, \theta), \alpha_i < 1, \alpha_j = 1)$. Such behavior saves costs for bank i with respect to regulatory tax. We define the cost of bank i at unilateral deviation as the *positive pivotal cost* of bank i, denoted PC_i^+. We define *negative pivotal cost* as $PC_i^-((\psi, \theta), \alpha_i = 1, \alpha_j < 1)$, i.e., when bank i unilaterally complies fully with regulatory policy (ψ, θ). In the latter case, bank i bears the full regulatory tax and cost of systemic risk for the case of a cost function with contagion effects. It is straightforward, then, to ascertain that

$$PC_i^+ \leq C_i((\psi, \theta), \alpha = \mathbf{1}) \leq PC_i^-. \tag{5}$$

The following is a permissive assumption.

Assumption 1. $PC_i^- - C_i((\psi, \theta), \alpha = \mathbf{1}) \geq C_i((\psi, \theta), \alpha = \mathbf{1}) - PC_i^+$.

The latter suggests that the additional cost of conforming to regulation when all others act strategically exceeds the cost saved by unilaterally circumventing it when all others conform. The assumption is satisfied for both cost function specifications in the model.

The average pivotal effect of bank i can be attributed by the *average pivotal cost* $APC_i = (PC_i^+ + PC_i^-)/2$ and gives information about the net effect of bankers' unilateral deviations. In the case of cost function with immunization (Eq. (1)) we anticipate that $C_i((\psi,\theta),\alpha = \mathbf{1}) = PC_i^-$, because the bank becomes fully immune to systemic risk and bears only the regulatory tax. Hence, it is $APC_i \leq C_i((\psi,\theta),\alpha = \mathbf{1})$.

Accordingly, we define *total average pivotal cost* as $TAPC = \sum_i APC_i$. Per Topkis (1998) (lemma 2.6.1. p49), we know that the sum of submodular functions is submodular and all properties of the (average) cost function are inherited by total (average) cost.

Proposition 2. *The PFA is bounded from above by* $\max_{\alpha^*}\left\{\frac{2TAPC-SOC}{SOC}\right\}$.

Proof. Appendix E.

Corollary 1. *Under Assumption 2, the PFA is bounded away from 1 and Nash equilibrium is always inefficient.*

Proof. Appendix F.

Proposition 2 indicates that all Nash equilibria are inefficient and that the upper bound indicates how detrimental it can be for banks to undertake off-balance-sheet activities. The higher the upper bound of PFA, the more susceptible the system is to banks' opportunism; nevertheless, we cannot make sure that the upper bound is actually attained. Under Eq. (2), this upper bound has important effects on the negative pivotal cost of banks (PC_i^-). The latter stands for the bank's cost to conform with regulatory policy when all competitors act strategically. The higher the differential in the equation of Assumption 2, the more costly conformance becomes. If we claim strict inequality in Assumption 2, the upper bound is always distant from 1.

Section 3 extends the model to include banks that willfully seek to destabilize the banking system.

3. THE GAME WITH "BYZANTINE" BANKERS

Economies can include willfully destabilizing agents - e.g., short-sellers who manipulate a collapse in stock prices or bond traders who pressure an issue

to trigger credit default swaps that they hold. Such trades are not always traceable. Dark pools that allow institutional investors to mask their activity from other market participants account for 14% of US stock trading volume (Buti et al., 2017). Dark pool trading platforms bypass and fragment open markets and disrupt market information.

Once trade masking appears, some bankers may seek to destabilize the financial system. We call these *malicious Byzantine bankers* after the Byzantine generals' problem in the computer network literature. We divide the game with *Byzantine Bankers* into two classes: a proper subset of profit maximizing bankers I^p, and malicious bankers I^m, with $I = I^p \cup I^m$. We assume that each banker in I^m aims to destabilize the system, seeking to maximize the difference $|PFA - 1|$. We anticipate that no Byzantine banker eventually opts for a positive α_i, whatever the cost to their balance sheets, as they profit with off-balance-sheet activities.

Definition 5. *The Byzantine regulatory arbitrage game is defined by the cost minimization game*

$$\Gamma = \{I, \{[0,1]\}_{i\in I}, \{C_i\}_{i\in I^p}, \{C_i\}_{i\in I^m}, (\psi, \theta)\}\}.$$

In this specification, overall social cost includes only costs incurred by profit maximizing bankers, $i \in I^p$. It is legitimate to exclude malicious Byzantine bankers from social cost because we assume they undermine social welfare. Malicious players have a destabilizing role in the economy; they favor an increasing social cost and operate as adversaries to profit-maximizing bankers. Therefore, excluding them from the overall social cost from a regulator's viewpoint is correct. Thus, $\bar{\mathbf{C}} = \sum_{i\in I^p} C_i((\psi, \theta, \alpha = \mathbf{1}|I^m)$. Accordingly, we define equilibrium in the game with Byzantine players, α^*, which we call *Byzantine Nash equilibrium*, and the overall cost to profit-maximizing bankers at equilibrium, $\mathbf{C}^* = \sum_{i\in I^p} C_i((\psi, \theta), \alpha^*|I^m)$.

We keep the information conditions of the game as abstract as possible. In the Byzantine agreement framework, it is customary to assume that we only know the presence of malicious players. Nevertheless, all the relevant information is available to define the price of financial anarchy in this context. Therefore, we modify the PFA to accommodate their presence. In this scenario, the worst-case equilibrium is given by a ratio that is far distant from 1. The following lemma provides a necessary condition for the latter.

Lemma 5. *The (profit maximizing) bank's cost at equilibrium will exceed the social optimum cost whenever*

$$\frac{\omega A_i}{RT_i} \geq \frac{n}{m+\sum_{j\in I^p}(1-\alpha_j)}.$$

Proof. Appendix G.

By lemma 1, the haircut (ωA_i) exceeds the regulatory tax (RT_i).That condition is likely satisfied when the number of banks (n) is relatively small, assuming so throughout the proof.

Definition 6 (Price of Byzantine Financial Anarchy)**.** *The Price of Byzantine financial anarchy (PBFA) is defined as the deviation from social optimum cost for worst- case equilibrium (in the equilibrium set). The ratio is*

$$PBFA(I^p;I^m) = \max_{\alpha^*} \frac{\mathbf{C}^*(I^p;I^m)}{\bar{\mathbf{C}}(I^p)}. \tag{6}$$

The PBFA captures the suboptimality of the worst-case Nash equilibrium in the extended version of the game with Byzantine bankers, and it is not much different from the PFA in practice. As claimed, we do not calculate the cost of Byzantine bankers in overall social cost. However, we consider their strategic influence on the cost of the remaining (profit-maximizing) bankers. Following Moscibroda et al. (2006) we define the *price of malice* (PoM) that conceptualizes the relative inefficiency for the original game.

Definition 7. *The Price of Malice (PoM) measures inefficiency in the system caused by Byzantine bankers and is given by the ratio*

$$PoM(I^m) = \frac{PBFA(I^p;I^m)}{PFA(I^p)}. \tag{7}$$

The PoM describes the degree of suboptimality resulting from Byzantine bankers. The lower the PoM is, the more tolerant the system is to the presence of malicious participants. Put differently, it measures how much damage is caused by the presence of Byzantine bankers.

Let us now see how detrimental the activity of Byzantine bankers can be to the financial system.

Proposition 3. *PBFA is bounded from above by the ratio*

$$\max_{\alpha^*} \left\{ \frac{2TAPC - SOC}{SOC} + m\Gamma \right\},$$

with $\Gamma = \sum_{I^p}(1-\alpha_j) \cdot \frac{\omega A_i}{n} > 0$.

Proof. Appendix H.

Interestingly, PBFA can be expressed as the decomposition of the original PFA attributed to the strategic inefficiency of profit-maximizing bankers and the component of financial risks originating with Byzantine bankers. This decomposition might help to assess differing regulatory constructs. For example, we can calculate how discouraging each proposed policy could be for profit-maximizing bankers and how immune the financial system becomes from the actions of Byzantine bankers. When policies primarily target the destabilizing role of Byzantine bankers, it might be more appropriate to use PoM as the indicator.

Corollary 2. *PoM in the Byzantine regulatory arbitrage game is*

$$PoM(I^m) = \frac{m\Gamma \cdot SOC}{2TAPC - SOC}.$$

Proof. The corollary follows directly from the definition. □

The PoM increases when the number of malicious bankers increases or the overall cost of fully complying with financial regulation rises.

4. DISCUSSION

This study addresses the problem caused by profit-maximizing bankers when they try to circumvent regulations and seek extra profit at the peril of system stability. The regulatory arbitrage game is an abstract but powerful framework for addressing banks' strategic considerations. Banks have the opportunity to increase their leverage and amplify profits. So long as these practices impose no cost on bankers and many financial instruments remain unregulated, malevolent motives remain. Drawing upon the "price of anarchy," we introduce the necessary theoretical underpinnings to capture social inefficiency caused by profit-maximizing bankers. To our knowledge, there is no other study that makes use of congestion games to contemplate the incentives of a bank's management in a regulated environment.

The price of financial anarchy is novel and can be used in three respects. First, to assess whether market regulations correct a vulnerability in financial systems; second, to calculate the critical PFA values that make the financial system fragile; third, to pursue a regulation that suppresses it below these thresholds. In a broad sense, financial fragility conceptualizes triggering a financial crisis by an exogenous (small) financial or economic shock. We emphasize that the more unregulated the financial system, the higher is the risk of triggering a crisis. Put differently, Byzantine bankers always seek to circumvent regulations to profit from financial turmoil, and that opportunity emerges in upturns and downturns.

Appendices

A. PROOF OF LEMMA 1

For the arbitrary strategy profile of banks α the cost function takes the form,

$$C_i = \alpha_i RT_i + (1-\alpha_i)^2 \frac{\omega A_i}{n} + (1-\alpha_i)\frac{\sum_{j\neq i}(1-\alpha_j)}{n}\omega \cdot A_i$$

We require for all $i \in I \setminus \{n+1\}$ to be $C_i > \bar{C}_i = RT_i$.

$$\begin{aligned}
\alpha_i RT_i + (1-\alpha_i)^2 \frac{\omega A_i}{n} + (1-\alpha_i)\frac{\sum_{j\neq i}(1-\alpha_j)}{n}\omega \cdot A_i &> RT_i \\
(1-\alpha_i)\frac{\omega A_i}{n}[(1-\alpha_i) + \sum_{j\neq i}(1-\alpha_j)] &> (1-\alpha_i)RT_i \\
\frac{\omega A_i}{n}\sum_{\forall j}(1-\alpha_j) &> RT_i \\
\frac{\omega A_i}{RT_i} &> \frac{n}{\sum_{\forall j \in I\setminus\{n+1\}}(1-\alpha_j)} \\
&> 1.
\end{aligned}$$

□

B. PROOF OF LEMMA 2

For an arbitrary i 's cost function in Eq. 1, the first derivative gives

$$\frac{\partial C_i}{\partial \alpha_i} = RT_i - 2(1-\alpha_i)\frac{\omega A_i}{n} - \sum_{j=1}^{\#S\setminus\{i\}} (1-\alpha_j)\frac{\omega \cdot A_i}{n}.$$

Now define the function $\hat{C}_i = (1-\alpha_i)^2 \frac{\omega A_i}{n}$ for which the first derivative is $\frac{\partial C_i}{\partial \alpha_i} = -2(1-\alpha_i)\frac{\omega A_i}{n}$.

Since the domain of α_i is a convex subset of reals and both $C_i, \hat{C}_i$ are quadratic, first-order conditions are sufficient for a minimum. Hence both achieve a minimum for some α_i. It is easily seen that both functions achieve minima for the same α_i. That is,

$$\arg\min_{\alpha_i} C_i(\alpha_i;\alpha_{-i}) = \arg\min_{\alpha_i} \hat{C}_i(\alpha_i;\alpha_{-i}).$$

Now we define the function,

$$P(\alpha) = \sum_{i}^{I\setminus n+1} \hat{C}_i = \sum_{i}^{I\setminus n+1} (1-a_i)^2 \frac{\omega \cdot A_i}{n}.$$

The function P is twice differentiable with respect to α_i and strictly convex. Differentiating,

$$\frac{\partial P}{\partial \alpha_i} = -2(1-\alpha_i)\frac{\omega A_i}{n} = \frac{\partial C_i}{\partial \alpha_i}. \tag{8}$$

it follows that

$$\arg\min_{\alpha_i} C_i(\alpha_i;\alpha_{-i}) = \arg\min_{\alpha_i} \hat{C}_i(\alpha_i;\alpha_{-i}) = \arg\min_{\alpha_i} P(\alpha_i;\alpha_{-i}),$$

which makes function $P(\alpha)$ an admissible best-response potential, i.e.,

$$\arg\min_{\alpha_i} C_i(\alpha) = \arg\min_{\alpha_i} P(\alpha).$$

□

C. PROOF OF LEMMA 3

By substituting (1) into the definition of decreasing differences (4), we have for the left side

$$C_i(\alpha_i,\alpha_j)-C_i(\alpha_i,\alpha_j')=\frac{(1-\alpha_i)\omega A_i}{n}\left(\sum_{j\in S\setminus\{i\}}(1-\alpha_j)-\sum_{j\in S\setminus\{i\}}(1-\alpha_j')\right).$$

Similarly, for the right side we have

$$C_i(\alpha_i',\alpha_j)-C_i(\alpha_i',\alpha_j')=\frac{(1-\alpha_i')\omega A_i}{n}\left(\sum_{j\in S\setminus\{i\}}(1-\alpha_j)-\sum_{j\in S\setminus\{i\}}(1-\alpha_j')\right).$$

For $\alpha_i\geq\alpha_i'$ it is always

$$\frac{(1-\alpha_i)\omega A_i}{n}\leq\frac{(1-\alpha_i')\omega A_i}{n}.$$

Hence $C_i(\alpha_i,\alpha_j)-C_i(\alpha_i,\alpha_j')\leq C_i(\alpha_i',\alpha_j)-C_i(\alpha_i',\alpha_j')$. □

D. PROOF OF LEMMA 4

We prove the lemma *mutatis mutandis* following the proof of Lemma 3.

For the left side,

$$\begin{aligned}C_i(\alpha_i,\alpha_j)-C_i(\alpha_i,\alpha_j')&=\alpha_i RT_i+\frac{(1-\alpha_i)}{n}\omega\cdot A_i+\frac{\sum_{j\neq i}(1-\alpha_j)}{n}\omega\cdot A_i\\&\quad-\alpha_i RT_i-\frac{(1-\alpha_i)}{n}\omega\cdot A_i-\frac{\sum_{j\neq i}(1-\alpha_j')}{n}\omega\cdot A_i\\&=\frac{\omega A_i}{n}\left(\sum_{j\in S\setminus\{i\}}(1-\alpha_j)-\sum_{j\in S\setminus\{i\}}(1-\alpha_j')\right).\end{aligned}$$

Similarly, for the right side we have

$$\begin{aligned}C_i(\alpha_i',\alpha_j)-C_i(\alpha_i',\alpha_j')&=\alpha_i' RT_i+\frac{(1-\alpha_i')}{n}\omega\cdot A_i+\frac{\sum_{j\neq i}(1-\alpha_j)}{n}\omega\cdot A_i\\&\quad-\alpha_i' RT_i-\frac{(1-\alpha_i')}{n}\omega\cdot A_i-\frac{\sum_{j\neq i}(1-\alpha_j')}{n}\omega\cdot A_i\\&=\frac{\omega A_i}{n}\left(\sum_{j\in S\setminus\{i\}}(1-\alpha_j)-\sum_{j\in S\setminus\{i\}}(1-\alpha_j')\right).\end{aligned}$$

Evidently, $C_i(\alpha_i, \alpha_j) - C_i(\alpha_i, \alpha'_j) = C_i(\alpha'_i, \alpha_j) - C_i(\alpha'_i, \alpha'_j)$. □

E. PROOF OF PROPOSITION 2

Per Topkis (1998)(Lemma 2.6.1, p. 49), we know a sum of submodular functions is submodular. Therefore, collective cost function $\mathbf{C} = \sum_i C_i$ is submodular. Consider the following profiles: $\bar{\alpha} = (1, \mathbf{1})$ and $\alpha^* = (\alpha_i^*, \alpha_{\mathbf{j}}^*)$ with $\bar{\alpha}$ being the profile of all bankers complying fully with regulation and corresponding to the optimal solution with regard to FRA, and α^* a Nash equilibrium.

By the property of decreasing differences,

$$
\begin{aligned}
C_i(\bar{\alpha}_i, \bar{\alpha}_j) - C_i(\bar{\alpha}_i, \alpha_j^*) &\leq C_i(\alpha_i^*, \bar{\alpha}_j) - C_i(\alpha_i^*, \alpha_j^*) \\
C_i(\bar{\alpha}_i, \bar{\alpha}_j) + C_i(\alpha_i^*, \alpha_j^*) &\leq C_i(\alpha_i^*, \bar{\alpha}_j) + C_i(\bar{\alpha}_i, \alpha_j^*) \\
C_i(\bar{\alpha}_i, \bar{\alpha}_j) + C_i(\alpha_i^*, \alpha_j^*) &\leq PC_i^+ + PC_i^- = 2APC_i \\
1 + \frac{C_i(\alpha_i^*, \alpha_j^*)}{C_i(\bar{\alpha}_i, \bar{\alpha}_j)} &\leq \frac{2APC_i}{C_i(\bar{\alpha}_i, \bar{\alpha}_j)} \\
\frac{C_i(\alpha_i^*, \alpha_j^*)}{C_i(\bar{\alpha}_i, \bar{\alpha}_j)} &\leq \frac{2APC_i - C_i(\bar{\alpha}_i, \bar{\alpha}_j)}{C_i(\bar{\alpha}_i, \bar{\alpha}_j)}.
\end{aligned}
$$

Applying summation by parts for all banks,

$$
\begin{aligned}
max_{\alpha^*} \frac{\mathbf{C}(\alpha_i^*, \alpha_j^*)}{\mathbf{C}(\bar{\alpha}_i, \bar{\alpha}_j)} &\leq max_{\alpha^*} \frac{\sum_n 2APC_i - SOC}{SOC} \\
PFA &\leq \max_{\alpha^*} \left\{ \frac{2TAPC - SOC}{SOC} \right\}.
\end{aligned}
$$

□

F. PROOF OF COROLLARY 1

Viz. the proof of Proposition 2 we ask

$$
\frac{2APC_i - C_i(\bar{\alpha}_i, \bar{\alpha}_j)}{C_i(\bar{\alpha}_i, \bar{\alpha}_j)} \geq 1,
$$

or

$$\begin{aligned} 2APC_i - C_i(\bar{\alpha}_i, \bar{\alpha}_j) &\geq C_i(\bar{\alpha}_i, \bar{\alpha}_j) \\ PC_i^+ + PC_i^- - C_i(\bar{\alpha}_i, \bar{\alpha}_j) &\geq C_i(\bar{\alpha}_i, \bar{\alpha}_j) \\ PC_i^- - C_i(\bar{\alpha}_i, \bar{\alpha}_j) &\geq C_i(\bar{\alpha}_i, \bar{\alpha}_j) - PC_i^+. \end{aligned}$$

Under Assumption 2, Corollary 1 is always true. □

G. PROOF OF LEMMA 5

Denote the cost function of bank i in the game with Byzantine bankers by $C_i^B(\alpha_i, \alpha_j; \alpha_{\mathbf{m}} = \mathbf{0})$, where vector $\alpha_{\mathbf{m}}$ denotes the strategy of malicious bankers. The cost function takes the form

$$\begin{aligned} C_i^B = \alpha_i RT_i + (1-\alpha_i)^2 \frac{\omega A_i}{n} + (1-\alpha_i) \frac{\sum_{j \in I^p \setminus \{i\}} (1-\alpha_j)}{n} \omega \cdot A_i \\ + (1-\alpha_i) \cdot m \cdot \frac{\omega \cdot A_i}{n}. \end{aligned}$$

The last term captures the cost effect of malicious bankers. Then it must be

$$C_i^B(\alpha_i, \alpha_j; \alpha_{\mathbf{m}} = \mathbf{0}) > C_i^B(\bar{\alpha}_i, \bar{\alpha}_j; \alpha_{\mathbf{m}} = \mathbf{0})$$

$$\begin{aligned} \alpha_i RT_i + (1-\alpha_i)^2 \frac{\omega A_i}{n} + (1-\alpha_i) \frac{\sum_{j \in I^p \setminus \{i\}} (1-\alpha_j)}{n} \omega \cdot A_i \\ + (1-\alpha_i) \cdot m \cdot \frac{\omega \cdot A_i}{n} > RT_i \end{aligned}$$

$$\begin{aligned} &(1-\alpha_i) \frac{\omega \cdot A_i}{n} [m + (1-\alpha_i) + \sum_{j \in I^p \setminus \{i\}} (1-\alpha_j)] > (1-\alpha_i) RT_i \\ &\frac{\omega \cdot A_i}{RT_i} > \frac{n}{m + (1-\alpha_i) + \sum_{j \in I^p \setminus \{i\}} (1-\alpha_j)} \\ &\frac{\omega \cdot A_i}{RT_i} > \frac{n}{m + \sum_{j \in I^p} (1-\alpha_j)}. \end{aligned}$$

□

H. PROOF OF PROPOSITION 3

We prove this viz. Proposition 2. We begin by estimating cost functions $C_i^B(\bar{\alpha}_i, \alpha_j^*; \alpha_{\mathbf{m}} = \mathbf{0})$ and $C_i^B(\alpha_i^*, \bar{\alpha}_j; \alpha_{\mathbf{m}} = \mathbf{0})$. It is easily verified that

$$C_i^B(\bar{\alpha}_i, \alpha_j^*; \alpha_{\mathbf{m}} = \mathbf{0}) = C_i(\bar{\alpha}_i, \alpha_j^*) = PC_i^-.$$

The bank becomes immune to systemic risk by fully complying with regulatory policy. In the same manner,

$$\begin{aligned} C_i^B(\alpha_i^*, \bar{\alpha}_j; \alpha_{\mathbf{m}} = \mathbf{0}) &= \alpha_i RT_i + (1-\alpha_i)^2 \frac{\omega A_i}{n} + (1-\alpha_i) \cdot m \cdot \frac{\omega \cdot A_i}{n} \\ &= C_i(\alpha_i^*, \bar{\alpha}_j) + (1-\alpha_i) \cdot m \cdot \frac{\omega \cdot A_i}{n} \\ &= PC_i^+ + (1-\alpha_i) \cdot m \cdot \frac{\omega \cdot A_i}{n}. \end{aligned}$$

Following the rationale of Proposition 2, we have

$$\begin{aligned} C_i^B(\bar{\alpha}_i, \bar{\alpha}_j; \alpha_{\mathbf{m}} = \mathbf{0}) + C_i^B(\alpha_i^*, \alpha_j^*; \alpha_{\mathbf{m}} = \mathbf{0}) &\leq C_i^B(\alpha_i^*, \bar{\alpha}_j; \alpha_{\mathbf{m}} = \mathbf{0}) \\ &\quad + C_i^B(\bar{\alpha}_i, \alpha_j^*; \alpha_{\mathbf{m}} = \mathbf{0}) \\ C_i^B(\bar{\alpha}_i, \bar{\alpha}_j; \alpha_{\mathbf{m}} = \mathbf{0}) + C_i^B(\alpha_i^*, \alpha_j^*; \alpha_{\mathbf{m}} = \mathbf{0}) &\leq PC_i^+ + (1-\alpha_i) \cdot m \cdot \frac{\omega \cdot A_i}{n} + PC_i^- \\ &\leq 2APC_i + (1-\alpha_i) \cdot m \cdot \frac{\omega \cdot A_i}{n} \\ 1 + \frac{C_i^B(\alpha_i^*, \alpha_j^*; \alpha_{\mathbf{m}} = \mathbf{0})}{C_i^B(\bar{\alpha}_i, \bar{\alpha}_j; \alpha_{\mathbf{m}} = \mathbf{0})} &\leq \frac{2APC_i + (1-\alpha_i) \cdot m \cdot \frac{\omega \cdot A_i}{n}}{C_i^B(\bar{\alpha}_i, \bar{\alpha}_j; \alpha_{\mathbf{m}} = \mathbf{0})} \\ \frac{C_i^B(\alpha_i^*, \alpha_j^*; \alpha_{\mathbf{m}} = \mathbf{0})}{C_i^B(\bar{\alpha}_i, \bar{\alpha}_j; \alpha_{\mathbf{m}} = \mathbf{0})} &\leq \frac{2APC_i - C_i^B(\bar{\alpha}_i, \bar{\alpha}_j; \alpha_{\mathbf{m}} = \mathbf{0})}{C_i^B(\bar{\alpha}_i, \bar{\alpha}_j; \alpha_{\mathbf{m}} = \mathbf{0})} \\ &\quad + \frac{(1-\alpha_i) \cdot m \cdot \frac{\omega \cdot A_i}{n}}{C_i^B(\bar{\alpha}_i, \bar{\alpha}_j; \alpha_{\mathbf{m}} = \mathbf{0})}. \end{aligned}$$

Applying summation by parts for all banks,

$$\begin{aligned} max_{\alpha^*} \frac{C^B(\alpha_i^*, \alpha_j^*; \alpha_{\mathbf{m}} = \mathbf{0})}{C^B(\bar{\alpha}_i, \bar{\alpha}_j; \alpha_{\mathbf{m}} = \mathbf{0})} &\leq max_{\alpha^*} \frac{\sum_n 2APC_i - SOC}{SOC} \\ PBFA &\leq \max_{\alpha^*} \left\{ \frac{2TAPC - SOC}{SOC} + \frac{\sum_{I^p} (1-\alpha_j) \cdot \frac{\omega A_i}{n}}{SOC} \right\}. \end{aligned}$$

□

References

Acharya, V., & Richardson, M. (2009). Causes of the financial crisis. *Critical Review*, *21*(2-3), 195–210.

Baer, H., & Pavel, C. (1988, March). Does Regulation Drive Innovation? *Economic Perspectives, Federal Reserve Bank of Chicago*, *66*, 3–15.

Breuer, P. (2002). Measuring Off-Balance-Sheet Leverage. *Journal of Banking & Finance*, *26*, 223–242.

Buti, S., Rindi, B., & Werner, I. M. (2017). Dark pool trading strategies, market quality and welfare. *Journal of Financial Economics*, *124*(2), 244–265.

Dubey, P. (1986). Inefficiency of Nash equilibria. *Mathematics of Operations Research*, *1*(11), 1–8.

GTen. (2001). *Report on consolidation in the financial sector* (Tech. Rep.). Group of Ten.

Jones, D. (2000). Emerging problems with the Basel Capital Accord: Regulatory capital arbitrage and related issues. *Journal of Banking and Finance*, *24*(1-2), 35–58.

Juarez, R. (2006, December). The worst absolute surplus loss in the problem of commons: random priority versus average cost. *Economic Theory*, *34*(1), 69–84.

Koutsoupias, E., & Papadimitriou, C. (2009, May). Worst-case equilibria. *Computer Science Review*, *3*(2), 65–69.

Lamport, L., Shostak, R., & Pease, M. (2019). The Byzantine generals problem. In *Concurrency: The Works of Leslie Lamport* (pp. 203–226).

Monderer, D., & Shapley, L. (1996, May). Potential Games. *Games and Economic Behavior*, *14*(1), 124–143.

Moscibroda, T., Schmid, S., & Wattenhofer, R. (2006). When selfish meets evil: Byzantine players in a virus inoculation game. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on Principles of Distributed Computing* (pp. 35–44). ACM.

Moulin, H. (2007, August). The price of anarchy of serial, average and incremental cost sharing. *Economic Theory*, *36*(3), 379–405.

Pavel, C., & Phillis, D. (1987, May). Why Commercial Banks Sell Loans: An empirical Analysis. *Economic Perspectives, Federal Reserve Bank of Chicago*, 3–14.

Roughgarden, T. (2006). Potential functions and the inefficiency of equilibria. *Proceedings of the International Congress of Mathematicians, Madrid, Spain*.

Roughgarden, T. (2012). The price of anarchy in games of incomplete information. *Proceedings of the 13th ACM Conference on Electronic Commerce*, *1*(212).

Topkis, D. (1998). *Supermodularity and complementarity*. Princeton University Press.

Vetta, A. (2002). Nash equilibria in competitive societies, with applications to facility location, traffic routing and auctions. *Foundations of Computer Science, 2002. Proceedings of the 43rd Symposium on the Foundations of Computer Science*.

Voorneveld, M. (2000, March). Best-response potential games. *Economics Letters*, *66*(3), 289–295.

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 `http://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.