Kidney Exchange with Multiple Donors*

Short Cycle Mechanisms

Anand Siththaranjan†

May 15, 2025

Abstract

This paper studies the problem of finding efficient, strategy-proof, and individually rational mechanism for kidney paired exchange under cycle-size restrictions. We allow patients to have multiple donors, where their private information is their strict preference over the donor used. We show that no desirable pairwise mechanism exists, but under mild conditions constructively show that such a two-and-three-way mechanism does exist. Our results leverage the structure of blood type compatibility to overcome classic impossibility results due to cycle-size restrictions, and we provide intuition to this end. When considering the number of transplants made, imposing strategy-proofness while allowing multiple donors introduces a tradeoff when compared to a perfect information single-donor baseline. We show through simulation using US population data that nearly always weak relative increases in the number of transplants, and with sufficiently many patients with more than one donor this can range from approximately 5 to 20%.

1 Introduction

One of the major successes of modern market design was the innovation of kidney paired exchange (KPE). When suffering from kidney failure, the most effective medical treatment is transplantation of a living kidney from a healthy donor. However, even when patients

*We thank Haluk Ergin for his support throughout this project.

†University of California, Berkeley, Department of Electrical Engineering and Computer Science, Berkeley, 2594 Hearst Ave, Berkeley, CA 94720, USA. E-mail: anandsranjan@berkeley.edu

have a willing donor, they may not be able to donate due to biological compatibility restrictions. KPE emerged as a solution to this problem by identifying sets of patient-donor pairs where patients could exchange their willing donors so that all patients can receive a compatible donation from a donor different than their own (Roth et al., 2004).

From the perspective of a patient, this requires them to identify potential donors of their own first. In general, a patient may have multiple donors to choose from, and may in fact have preferences over them. These preferences could reflect objective medical features such as the risk of donation for different donors, or other subjective features such as personal relationships, familial obligations, and more. For example, some donors may have a family whereas others don't, and in the case where they are the sole provider for their family then it may be difficult to take time off from work due to post-operative care. As such, real-world implementations of KPE mechanisms allow agents to bring multiple donors. However, these mechanisms tend to not solicit preferences, which can lead to strategizing on the part of patients. For example, patients may choose to not bring certain less preferred donors to the mechanism before attempting to be matched using a preferred donor first.

From the perspective of a designer this behaviour is not ideal. We would prefer patients to bring all their available donors to the exchange rather than gradually over time so that we can generate efficient outcomes. Achieving efficiency also requires patients to be able to express preferences over their donors. This motivates our work in trying to design an efficient, strategy-proof and individually rational mechanism when donors can be strictly ranked. Of key importance in the practice of KPE is ensuring the small cycles are used, as transplantation surgeries for KPE tend to be simultaneous. This paper aims to bridge a gap in the current literature on market design for KPE, which generally does not allow for general preferences over a patient's set of donors. First we consider the case where an exchange must be pairwise, finding an impossibility result in this environment. Pairwise exchanges are commonly one of the first studied settings for new organ exchange mechanisms with cycle constraints (Roth et al., 2005; Ergin et al., 2020, 2017), and this impossibility result contrasts with positive results when agents are indifferent over all their donors (Roth et al., 2005). In and of itself, this may not be surprising considering the literature on house exchange with cycle constraints (Balbuzanov, 2020). However the results of the latter do not directly imply results in our environment. In particular, when allowing for two-and-three cycle exchanges, we find a positive result under mild assumptions. Throughout the paper, we will develop intuition for how the structure of our problem places is amenable to cycle constraints.

A key metric by which KPE designers evaluate solutions on is the number of matches made. Imposing strategyproofness will generally lead to the first-best outcome not being implementable, whereas allowing multiple donors provide more opportunities to be matched over a single-donor baseline. Through simulations on US population data for various probabilities that a patient may have multiple donors, we compare of our mechanism with this baseline that has perfect information. We generally find weak improvements in terms of the relative increase in the number of matches made, and with moderate to high probabilities this ranges from approximately 5 to 20%. We conclude with a discussion of additional properties of our mechanism, and comments on the environment we study.

2 Model

Let $\mathcal{B} = \{A, B, AB, O\}$ be the set of blood types. A donor with blood type b can donate to a patient with blood type b' if $b \leq_{\mathcal{B}} b'$, where $\leq_{\mathcal{B}}$ is a reflexive, transitive partial order[1] given by the following:

1. O is a universal donor: $O \leq_{\mathcal{B}} b$ for all $b \in \mathcal{B}$
2. AB is a universal recipient: $b \leq_{\mathcal{B}} AB$ for all $b \in \mathcal{B}$
3. A and B are incompatible: $A \not\leq_{\mathcal{B}} B$ and $B \not\leq_{\mathcal{B}} A$
4. AB cannot donate to other types: $AB \not\leq_{B} b$ for $b \in \mathcal{B} \setminus \{AB\}$
5. O cannot receiver from other types: $b \not\leq_{B} O$ for $b \in \mathcal{B} \setminus \{O\}$

An agent is a patient-donor tuple $i \in \mathcal{I}$, where we denote $\tau_i = P - D_1, D_2, D_3, D_4, D_5 \in \mathcal{T} = \cup_{n=1}^{5} \mathcal{B} \times (\mathcal{B} \cup \{\emptyset\})^n$ to be the type of an agent. Let $\mathcal{D}_i = \{D_1, D_2, D_3, D_4, D_5\} \cap \mathcal{B}$. The type τ_i contains the following observable and private information:

1. P is the observable blood type of a patient
2. $\mathcal{D}_i \cup \{\emptyset\}$ are blood types of donors or the outside option $\emptyset$, which is referring to using no donor and thus not participating in any exchange.

[1]The biological characterization is as follows. A and B are antigens, and your blood type specifies if you have either antigen. For example, AB means you have both, whereas O means you have neither. A donor can donate to a recipient if whenever a recipient is missing an antigen, then the donor is also missing the antigen.

3. The unordered set $\mathcal{D}_i$ is the observable set of donor blood types.

4. The ordered set composed of $\mathcal{D}_i \cup \{\emptyset\}$ is the private preference ordering $\succ_i$ on donors and the outside option $\emptyset$, i.e. $D_1 \succ_i D_2 \succ_i D_3 \succ_i D_4 \succ_i D_5$.

We can assume that patients list up to 4 donors with no repeated blood types, any donor dis-preferred to the outside option will nevertheless be listed, and patients may be blood type compatible with their donors. We will find these assumptions to be without loss in our mechanism. We will also use the notation $\mathcal{I}_{P-D_1*}$ to refer the set of agents with blood type P for their patient, D_1 for their most-preferred donor, and $*$ means that they may have further donors. Furthermore, $n_{P-D_1*} = |\mathcal{I}_{P-D_1*}|$. Similarly we define $\mathcal{I}_{P-D_1,D_2*}$ where D_2 is the agent's second favourite donor, and so forth. Using the same notation, we succinctly write an agent's type $\tau_i = P - D_1*$ to emphasize that the D_1 donor, their most-preferred donor, is most relevant.

A **paired exchange environment** is $\mathcal{E} = (\mathcal{I}, \mathcal{T})$. Given such an environment, we use the notation $i \to_b j$ for $i, j \in \mathcal{I}$ and $b \in \mathcal{B}$ to mean that i **donates to** j **using a** b **donor**. We only consider feasible donations, that is a donation $i \to_b j$ such that

1. i has a donor with blood type b, i.e. $b \in \mathcal{D}_i$

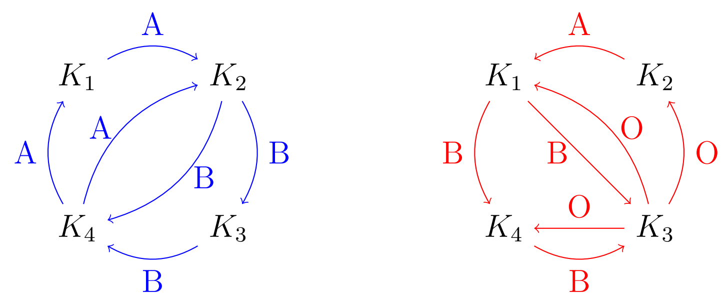
2. b can donate to the patient in j, i.e. $b \leq_{\mathcal{B}} P_j$

An environment $\mathcal{E}$ induces an edge-labeled directed multi-graph we refer to as a **compatibility graph**: $G_{\mathcal{E}} = (\mathcal{I}, F)$, where the set of labeled edges $F \subseteq \mathcal{I} \times \mathcal{I} \times \mathcal{B}$ is such that $(i, j, b) \in F$ if and only if $i \to_b j$. An exchange $E = \{(i_1^l \to \cdots \to i_{k_l}^l)\}_{l \in L}$ is a set of vertex-disjoint cycles indexed by L, where a single cycle $l \in L$ is $i_1^l \to \cdots \to i_{k_l}^l$ made up of distinct agents. The edge used is inferred based on the preferences of the agent from which it is outgoing, whereby the $i_k^l \to i_{k+1}^l$ using their most preferred donor b such that $i_k^l \to_b i_{k+1}^l$. An n-exchange is an exchange E such that for all cycles $l \in L$, we have that $k_l \leq n$. We refer to the set of exchanges by $\mathbb{E}$, and the set of n-exchanges by $\mathbb{E}^n$. In the special case where $n = 2$, we refer to an exchange as a **matching** and define $\mathbb{M} = \mathbb{E}^2$. We use the notation $E(i)$ to refer to the blood type of the donor used by i in E, and $\emptyset$ otherwise. Formally, $E(i) = b$ if there exists $j \in \mathcal{I}$ such that $(i, j, b) \in E$, and $\emptyset$ otherwise. Note that this is well-defined as exchanges are vertex disjoint, hence there is a unique donor used and agent that is donated to.

Example 1 (Two Donor Example)**.** Consider the environment $\mathcal{E}$ given by Figure 1a, with compatibility graph for D_1 and D_2 edges given by Figures 1b and 1c respectively. The

$\mathcal{I}$	P	D_1	D_2
K_1	A	A	B
K_2	A	B	A
K_3	B	B	O
K_4	B	A	B

(a) Environment $\mathcal{E}$



(b) $G_{\mathcal{E}}$ with Donor 1 (c) $G_{\mathcal{E}}$ with Donor 2

Figure 1: An example environment with two donors.

set of agents can be written by $\mathcal{I} = \{K_i\}_{i=1}^4$ where $K_1 = A - A, B$, $K_2 = A - B, A$, $K_3 = B - B, O$, and $K_4 = B - A, B$. △

2.1 Mechanisms

Given an environment $\mathcal{E}$ where $n = |\mathcal{I}|$, an **exchange mechanism** is $\phi : \mathcal{T}^n \to \mathbb{E}$. ϕ is a **pairwise mechanism** if $\phi[\mathcal{T}^n] \subseteq \mathbb{M}$, and a **two-and-three-cycle mechanism** if $\phi[\mathcal{T}^n] \subseteq \mathbb{E}^3$. Let $\bar{\mathbb{E}}$ be the range of ϕ.

Consider some $\tau \in \mathcal{T}^n$ and $E = \phi[\tau]$. We say that E is

1. **individually rational** (IR) if for all $i \in \mathcal{I}$, $E(i) \succeq_i \emptyset$.

2. **Pareto efficient** (PE) if there does not exists $E' \in \bar{\mathbb{E}}$ such that for all $j \in \mathcal{I}$, $E'(j) \succeq_j E(j)$, and there exists some $i \in \mathcal{I}$ such that $E'(i) \succ E(i)$.

A mechanism ϕ is IR and PE if for every $\tau \in \mathcal{T}^n$, $E = \phi[\tau]$ is IR and PE. A **deviation** for agent i from $\tau_i = P_i - D_1, \ldots, D_5$ is $\tau_i' = P_i - D_{k_1}, \ldots, D_{k_5}$, where $\{k_i\}_{i=1}^5$ is a permutation of $\{1, \ldots, 5\}$. A mechanism is **strategy-proof** (SP) if for all $\tau \in \mathcal{T}^n$, $i \in \mathcal{I}$ and deviations τ_i' from τ_i, we have that $\phi[\tau](i) \succeq_i \phi[\tau_i', \tau_{-i}](i)$. We say a mechanism is **desirable** if it is PE, IR, and SP.

3 Pairwise Mechanisms

We begin with studying the most minimal setting of short cycle mechanisms - that is pairwise exchange mechanisms. We ask whether there exists a desirable pairwise exchange mechanism for all environments. The following result shows that this does not hold.

Proposition 1. *There exists an environment where there is no desirable pairwise exchange mechanism.*

The structure of this proof identifies a counter-example with three agents whose patients and donors have A and O blood types only[2]. By leveraging all three criteria of desirability - strategy-proofness, Pareto efficiency, and individual rationality - we can consider different feasible matches at different type profiles and find a contradiction should all these criteria simultaneously hold.

Importance of Compatibility Structure. An important feature of this problem which has been ignored in past work is on incorporating the structure of biological compatibility. In particular, Gilon et al. (2019) has previously studied kidney exchange with multiple donors and the existence of pairwise mechanisms. By example, they aim to show the impossibility of a desirable mechanism under a pairwise cycle constraint. However, as we show below, their example is not possible under the standard blood type compatibility model. In particular, the structure of blood type compatibility restricts the set of feasible trades. Let $G = (\mathcal{I}, H)$ be a candidate compatibility digraph where $H \subseteq \mathcal{I} \times \mathcal{I}$. Note that this compatibility graph does not specify the blood type of the donor used in a directed edge. We say that G is **rationalizable** if there exists an environment $\mathcal{E}$ such that for $G_{\mathcal{E}} = (\mathcal{I}, F)$, we have that $H = \mathbf{proj}(F) = \{(i,j) \in \mathcal{I} \times \mathcal{I} | \exists b \in \mathcal{B},\ (i,j,b) \in F\}$. In other words, there exists a choice of blood types for patients and donors that make these matches as exactly the feasible set of matches.

Proposition 2. *The set of feasible pairwise matches given in Figure 1 of Gilon et al. (2019) is not rationalizable.*[3]

This result is illustrative about the importance of accounting for the blood-type compatibility structure. Though they find an impossibility result conjectured to be in the same setting as ours, it does not provide an underlying blood type structure to rationalize their example. We will see in the subsequent section that accounting for blood type compatibility is essential for a positive result. In light of this, we can view their result as a comment on the existence of desirable mechanisms when any trade is feasible and preferences over the outside option can be arbitrary, which may be useful in other areas of interest.

[2]Given that O and A are among the two most common blood types, this is not an edge case example.

[3]Figure 1 of Gilon et al. (2019) was used to prove the analogous negative result in their Theorem 3.

4 A Two-and-Three Cycle Mechanism

We now explore the existence of desirable mechanisms when cycles of size at most three can be used. We consider various assumptions either for simplicity in the design of the mechanism, or based in realistic assumptions on the distribution of agents.

Our first assumption simplifies the design of our mechanism by removing from consideration an uncommon blood type.

Assumption 1. *There are no patients with an AB blood type.*

As AB donors can only donate to AB patients, this assumption implies that it is without loss of generality that all agents only have non-AB donors, and have at most three donors. We provide further discussion of this assumption in Section 6. This next assumption is common in the literature, as stated in Roth et al. (2007), which is motivated by the intuition that there are many $O-A$ and $O-B$ pairs due to blood-type incompatibility between patient and donor[4] and the high proportion of O blood types[5].

Assumption 2 (Long-side of the Market)**.** *At least one of each type in $O-b$ for $b \in \{A, B\}$ is unmatched in any feasible matching.*

This final assumption is for simplicity of the mechanism, similar to that in Roth et al. (2007) except we preclude the possibility of no $b-b*$ agents:

Assumption 3. *There are at least two $O-O*$, $B-B*$, and $A-A*$.*

Let $\texttt{MaxMatch}(A)$ compute a maximum exchange of size at most 3 among all agents $A \subseteq \mathcal{I}$. Without loss of generality, we assume $n_{A-B*} \geq n_{B-A*}$, and our algorithm is analogous in the case where $n_{A-B*} < n_{B-A*}$. When we say that an agent i of type $X-YZ*$ **drop** their Y donor, we transform their type to be $X-Z*$ and eliminate their donor Y from being considered. Let Π be a priority order over $\mathcal{I}$, and $\Pi|_A$ be the same order restricted to $A \subseteq \mathcal{I}$. For a set of exchanges M and M', let the operation $M \leftarrow M'$ mean to add M' to M, and to remove all agents in M' from $\mathcal{I}$.

Consider the following mechanism: $M = \emptyset$

1. Match $\mathcal{I}_{B-B*}$ amongst each other: $M \leftarrow \texttt{MaxMatch}(\mathcal{I}_{B-B*})$.

[4]This leverages the idea that patients tend to enter the exchange after trying to use their donors but failing due to compatibility.

[5]Approximately 48% in the US.

2. Match $\mathcal{I}_{A-B*}$ and $\mathcal{I}_{B-A*}$: for $i \in \Pi_{\mathcal{I}_{A-B*}}$.

   (a) If $n_{B-A*} \neq 0$, then choose some $j \in \mathcal{I}_{B-A*}$ and do $M \leftarrow \{(i \to j)\}$.

   (b) Else, exit this for loop.

3. Serial Dictator-like procedure with agents $O - A*$ and $O - B*$: for $i \in \Pi_{\mathcal{I}_{O-A*} \cup \mathcal{I}_{O-B*}}$

   (a) If $n_{B-O*} = 0$ then for all $j \in \mathcal{I}_{A-B*}$, j drops their B donor.

   (b) If $i \in \mathcal{I}_{O-A*}$ and

      i. if $n_{A-B*} > 0$ and $n_{B-O*} > 0$, then for $j \in \max_{\Pi_{A-B*}} \mathcal{I}_{A-B*}$ and $k \in \mathcal{I}_{B-O*}$ do $M \leftarrow \{(i \to j \to k)\}$.

      ii. Else, if $n_{A-O*} > 0$, then for $k \in \mathcal{I}_{A-O*}$ do $M \leftarrow \{(i \to j)\}$.

      iii. Else, if $i \in \mathcal{I}_{O-A,B*}$ and $n_{B-O*} > 0$, then for $j \in \mathcal{I}_{B-O*}$ do $M \leftarrow \{(i \to j)\}$.

   (c) If $i \in \mathcal{I}_{O-B*}$ and

      i. if $n_{B-O*} > 0$, then for $k \in \mathcal{I}_{B-O*}$ do $M \leftarrow \{(i \to j)\}$.

      ii. Else, if $i \in \mathcal{I}_{O-B,A*}$ and $n_{A-O*} > 0$, then for $j \in \mathcal{I}_{A-O*}$ do $M \leftarrow \{(i \to j)\}$.

4. For $i \in \mathcal{I}_{\mathcal{I}_{O-A*} \cup \mathcal{I}_{O-B*}}$, i drops their A and B donors (if any).

5. Match $\mathcal{I}_{O-O*}$ amongst each other: $M \leftarrow \texttt{MaxMatch}(\mathcal{I}_{O-O*})$.

6. Match $\mathcal{I}_{A-A*}$ amongst each other: $M \leftarrow \texttt{MaxMatch}(\mathcal{I}_{A-A*})$.

For intuition, we informally describe the algorithm as follows:

1. Match all $B - B*$ agents amongst themselves, allowing them to have their best option

2. As there are more $A - B*$ than $B - A*$, use a priority order to let $A - B*$ be matched with $B - A*$. Both agents get their best option.

3. Let $O - A*$ and $O - B*$ take turns choosing agents to match with.

   (a) $O - A*$ prioritizes matching with $A - B*$ and $B - O*$, then $A - O*$. Once they've exhausted their opportunities to use A, if they have a B donor as their second preferred donor, they match with $B - O*$ if possible.

   (b) $O - B*$ prioritizes $B - O*$, then $A - O*$ if possible (and A is their second preferred donor). There are no $B - A*$ agents left for a three-way exchange with an $A - O*$.

(c) $A-B*$ keeps their top donor until they no longer have any opportunities to use them.

4. $O-A*$ and $O-B*$ have no more matching opportunities with their A and B donor(s) and thus drop them.

5. Match all $O-O*$ and $A-A$ agents amongst themselves.

We now state the main properties of this mechanism:

Theorem 1. *This mechanism is PE, IR, and SP.*

To prove this result, the main challenge lies in showing Pareto efficiency. The proof technique employs a contradiction argument, where we assume that there is an agent i_0 who can be strictly improved from our matching M in some other Pareto dominating matching M'. We then construct a sequence of agents $\{i_n\}_{n\in\mathbb{N}}$ in which we have that i_n *steals* from i_{n+1}. By this we mean i_n is matched with some agent in M' who was originally matched to i_{n+1}. To maintain efficiency, i_{n+1} must be re-matched in M' with an agent they at least weakly prefer to whom they were matched to in M. By ensuring that no agent in the sequence is stolen from twice, then no agent ever repeats. Furthermore, we need to ensure that there is always an agent stolen from, rather than taking an unmatched agent. This is simple in the case of pairwise exchanges, but more difficult when we consider three-way exchanges. Given we are able to do this, the sequence must contain infinitely many distinct agents, which is a contradiction to there being finitely many agents in our environment. To achieve this result, we characterize the cases of i_0 that could potentially be improved, and construct a sequence for that case. For example, it may be that they were not matched via their top donor, hence can be improved.

5 Simulations

We compare our mechanism with a random priority order to a baseline where agents only bring one donor. We see this baseline as a setting where agents strategize and bring their most preferred donor, and consider the outcome from the maximum 2-or-3 cycle exchange mechanism[6] under complete information of preferences. As most papers study a single donor environment, or do not account for general donor preferences, we aim to highlight the practicality of mechanisms that do account for these factors.

[6]Such a mechanism can be found in Roth et al. (2007).

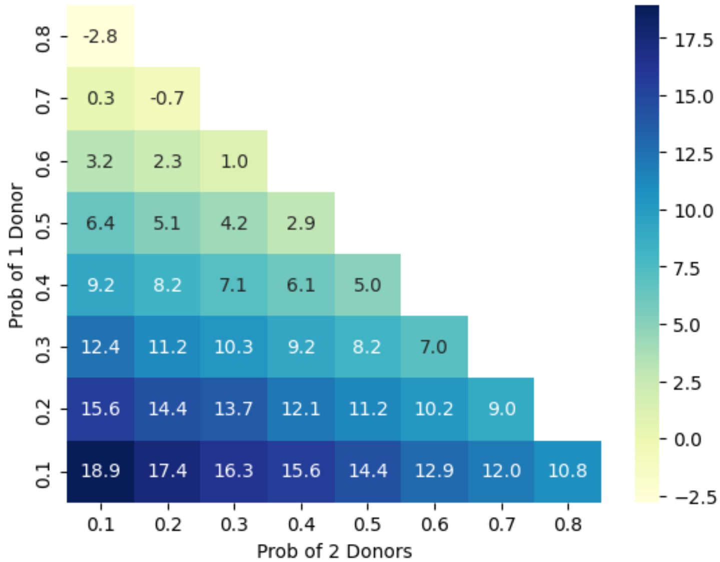


Figure 2: (%) Relative Increase for $n = 300$ given different probabilities of the number of donors. Averaged over 1000 random trials. The probability of 3 donors is given by $1 - \mathbb{P}(1 \text{ donor}) - \mathbb{P}(2 \text{ donors})$.

We randomly sample patients and donors from a common blood type distribution, and consider different probabilities that an agent may have one, two, or three donors. We only allow patients to have a top donor with the same blood type as them if they are tissue-type incompatible. We use the same probabilities as in Roth et al. (2007), where for each patient we sample a PRA level from $\{L, M, H\}$ and then randomly sampling with the relevant probabilities whether they are tissue-type compatible. Furthermore, we only consider distributions of agents that satisfy our assumptions in Section 4. Figure 2 shows the average relative increase in our mechanism compared to the baseline when there are 200 agents for different donor number probabilities. We average our results over 1000 random trials.

Improvements over the baseline are dependent on the number of agents with more than one donor. We can see in Figure 2 that there are approximately little, if any, losses from imposing strategyproof-ness, and relative increases from approximately 5 to 20% when at least 60% of agents have more than one available donor.

6 Discussion

How Many Donors Does One Need? We say that a mechanism is **k-donor equivalent** for $A \subseteq \mathcal{I}$ if for any pair $i \in A$, their outcome in the mechanism is the same than if they truncated their preference to their top k alternatives.[7]

Proposition 3. *Assume that $n_{A-B*} \geq n_{B-A*}$. The mechanism is*

1. 1-*donor equivalent for $\mathcal{I}_{B-*}$,*
2. 2-*donor equivalent for $\mathcal{I}_{A-*}$, and*
3. 3-*donor equivalent for $\mathcal{I}_{O-*}$.*

Proof. The claim can be seen by observing the worst case outcomes throughout the algorithm, noting that we proceed down an agent's donor list.

□

This result is useful in that certain patients need not spend more effort than necessary to find donors outside of their top k, where k is given by the above proposition.

TTC and Cycle Constraints. When there are no cycle constraints, variants of TTC that account for indifferences can easily be applied to our problem by the following. Let preferences over donation modes induce preferences over agents where agent i is strictly preferred to agent j by agent k if k can donate to i using a donation mode that is strictly better than any donation mode that k can use to donate to j. Thus, approaches that solve house exchange problems when there are indifferences, such as the Top Trading Absorbing Sets algorithm in Alcalde-Unzu and Molis (2011), can be used.

However, in our environment we require exchanges to be composed of two or three way cycles, that is cycles in an exchange are of length at most three. It is well known in the literature that imposing cycle constraints makes it such that there is generally no mechanism that is efficient, strategy-proof, and individually rational for house exchange. So why is it possible in this environment to find such a mechanism when there is a cycle constraint? The following result gives some intuition by showing that if there is a top trading cycle, then there is one that has length two (i.e. a 2-cycle):

[7]Formally, for a given preference $\succ_i$ over $\mathcal{M} \cup \{\emptyset\}$, denote the **$k$-truncated preference** $\succ_i^k$ such that there exists $m^l \in \mathcal{M}$ distinct where $m^1 \succ_i \cdots \succ_i m^{|\mathcal{M}|}$ and $m^1 \succ_i^k \cdots \succ_i^k m^k \succ_i^k \emptyset \succ_i^k m^{k+1}$. Then a mechanism ϕ is **k-donor equivalent** for A if for all preference profiles $\succ$ and for all agents $i \in A$, then $\phi(\succ) = \phi(\succ_i^k, \succ_{-i})$.

Proposition 4. *Consider a graph where all agents point to their favourite feasible agent. If there is a cycle, then there must be a 2-cycle.*

This result, and its proof, show that the structure of the compatibility relation and donor preferences induces a certain structure on preferences over agents that is somewhat compatible with the concept of a top trading cycle. Though this provides some intuition for why it is possible, note however that we cannot use the TTC algorithm itself due to indifferences, and some approaches that adapt TTC to allow for indifferences cannot clearly be adapted to leverage this 2-cycle existence property. For example, when cycles are executed in Alcalde-Unzu and Molis (2011), objects are only provisionally assigned and thus may be reassigned. As such we can't guarantee that all objects will remain "close" to their original owner. Further work should explore how to exploit such structure to show the existence of satisfactory mechanisms in general.

Number of blood types and cycle size. We might ask whether there is a connection between the number of blood types present and the maximum cycle size. For example, Roth et al. (2007) find that when there are n blood types that satisfy certain compatibility conditions, cycles of size at most n are sufficient to maximize the number of transplants amongst all exchanges with unbounded cycle size. Hence is it intuitive that we are able to find an efficient mechanism considering we only use three blood types? This observation does not straightforwardly hold as we find a negative result in the case of pairwise mechanism in Proposition 1 through a counter-example that only requires two blood types.

Assumption on AB patients. We now provide further justification for the assumption to not include AB patients in the design of our mechanism. Firstly, in the United States, AB blood types make up approximately 4% of the population. In general, it is a very rare blood type. Secondly, as AB patients can receive from any blood type, we can note that the probability that their top donor is also one they are tissue-type incompatible with is approximately 20%[8]. Hence they are unlikely to enter a paired exchange mechanism, and our assumption can be seen as implying that we do not consider a group of patients that make up less than 1% of the population. As such, the gains from incorporating agents with AB patients is likely to be minimal, and these patients are also those that tend to be easy to match.

[8]This can be calculated using the probabilities in Roth et al. (2007).

Preferences over Kidneys. Absent from our model are preferences over kidneys, which is commonly allowed for in the literature. As we primarily focus on incentives associated with donor preferences, we opt to exclude this aspect for simplicity and view this as an extension of models with 0-1 preferences for kidney exchange (Roth et al., 2005). We contend that this model still provides useful insights on the timing of matching agents and how this relates to preferences and blood types. Furthermore, from the perspective of surgeons, it has been noted in past work that they recommend indifference over living organs due to the similarity in outcomes (Roth et al., 2005; Yılmaz, 2011). For example, we can consider the age of a donor as a metric for health. Figure 4 in Terasaki et al. (1995) shows similar graft survival rates over a three year period for donors under the age of 50. When preferences are observable, as sometimes assumed in the literature due to the preferences for healthier kidneys, we may be able to incorporate them using a similar approach in Ergin et al. (2020).

7 Conclusion

In this work we consider the existence of desirable - that is Pareto efficient, individually rational, and strategy-proof - mechanisms for paired kidney exchange with multiple donors when cycles have size constraints and agents have preferences over which donor of theirs is used. We find negative results in the case of pairwise mechanisms, and positive results when allowing cycles of size at most three under reasonable assumptions. Our result is surprising in light of known results on cycle constraints and recent work on the same topic (Gilon et al., 2019), and our progress emphasizes the importance of leveraging biological compatibility in these problems. By further testing this mechanism in simulation using US population data, we can see reasonable improvements over single-donor baselines that motivate the practical viability of multi-donor mechanisms.

References

Alcalde-Unzu, Jorge and Elena Molis (2011), “Exchange of indivisible goods and indifferences: The top trading absorbing sets mechanisms.” *Games and Economic Behavior*, 73(1), 1–16. 11, 12

Balbuzanov, Ivan (2020), “Short trading cycles: Paired kidney exchange with strict ordinal preferences.” *Mathematical Social Sciences*, 104, 78–87. 2

Ergin, Haluk, Tayfun Sönmez, and M Utku Ünver (2017), “Dual-donor organ exchange.” *Econometrica*, 85(5), 1645–1671. 2

Ergin, Haluk, Tayfun Sönmez, and M Utku Ünver (2020), “Efficient and incentive-compatible liver exchange.” *Econometrica*, 88(3), 965–1005. 2, 13

Gilon, Oren, Tal Gilon, and Assaf Romm (2019), “Kidney exchange with multiple donors.” *Available at SSRN 3451470*. 6, 13, 17

Roth, Alvin E, Tayfun Sönmez, and M Utku Ünver (2004), “Kidney exchange.” *The Quarterly journal of economics*, 119(2), 457–488. 2

Roth, Alvin E, Tayfun Sönmez, and M Utku Ünver (2007), “Efficient kidney exchange: Coincidence of wants in markets with compatibility-based preferences.” *American Economic Review*, 97(3), 828–851. 7, 9, 10, 12

Roth, Alvin E., Tayfun Sönmez, and M. Utku Ünver (2005), “Pairwise kidney exchange.” *Journal of Economic Theory*, 125(2), 151–188, URL `https://www.sciencedirect.com/science/article/pii/S0022053105001055`. 2, 13

Terasaki, Paul I, J Michael Cecka, David W Gjertson, and Steven Takemoto (1995), “High survival rates of kidney transplants from spousal and living unrelated donors.” *New England Journal of Medicine*, 333(6), 333–336. 13

Yılmaz, Özgür (2011), “Kidney exchange: An egalitarian mechanism.” *Journal of Economic Theory*, 146(2), 592–618. 13

A Proofs

A.1 Proof of Proposition 1

Proof. Consider the following agents, that is patients with donors (that may or may not be feasible):

- $\{i, j, k\}$ each with an O patient and $\{O, A\}$ donors.
- l with an A patient and an O donor.

Observe that for any of $\{i, j, k\}$ to match with each other, it must be through their O donor, and to match with l can be through either donor. However to use their A donor, they must match with l. We will define a match by the agents participating in it, with their favourite donor used implicitly. For example (i, j) means that the donate to each other via their O donors. We say a match is valid if it is efficient and IR given the preferences considered.

Assume for contradiction that there is a desirable mechanism. Fix l's preference as $O \succ \emptyset$. Now consider the following preference profile $\succ_1$ for $\{i, j, k\}$:

$$i : A \succ \emptyset \succ O$$
$$j : A \succ \emptyset \succ O$$
$$k : A \succ \emptyset \succ O$$

Without loss of generality, let the valid allocation in the desirable mechanism match (i, l) together. Consider the following profile $\succ_2$:

$$i : O \succ A \succ \emptyset$$
$$j : A \succ \emptyset \succ O$$
$$k : A \succ \emptyset \succ O$$

Since (i, j) and (i, k) are not IR, then to maintain strategy-proofness we must have (i, l) match. Now consider the following $\succ_3$:

$$i : O \succ A \succ \emptyset$$
$$j : A \succ O \succ \emptyset$$
$$k : A \succ \emptyset \succ O$$

The valid matches are $\{(i,j),(k,l)\}$ and (j,l). However choosing the latter would violate strategy-proofness as there would be a profitable deviation for j to misreport in preference profile $\succ_3$ into $\succ_2$ where they go from being unmatched to being part of an individually rational match. Hence let the match here be $\{(i,j),(k,l)\}$. Now consider the following $\succ_4$:

$$i : O \succ \emptyset \succ A$$
$$j : A \succ O \succ \emptyset$$
$$k : A \succ \emptyset \succ O$$

Note that (j,l) is valid but not strategy-proof, as otherwise i will misreport from $\succ_4$ to $\succ_3$ and get from being unmatched to matched with j. Thus the only valid match is $\{(i,j),(k,l)\}$. Now consider the following $\succ_5$:

$$i : O \succ \emptyset \succ A$$
$$j : A \succ \emptyset \succ O$$
$$k : A \succ \emptyset \succ O$$

Note that (j,l) is valid but again not strategy-proof, as otherwise j in $\succ_4$ would misreport in $\succ_5$ to $\succ_4$ and go from (i,j) to (j,l), which is a strictly better outcome for them. Hence the only valid outcome is (k,l). Now consider the following $\succ_6$:

$$i : O \succ \emptyset \succ A$$
$$j : A \succ \emptyset \succ O$$
$$k : A \succ O \succ \emptyset$$

Observe if (k,l) is not the valid match chosen here, then k in $\succ_6$ will misreport to $\succ_5$ to get this match and thus strictly improve. Hence (k,l) is the outcome here. Now consider the following $\succ_7$:

$$i : O \succ A \succ \emptyset$$
$$j : A \succ \emptyset \succ O$$
$$k : A \succ O \succ \emptyset$$

There are two valid outcomes here: $\{(i,k),(j,l)\}$ and (k,l). Note that the former would not be strategy-proof, as otherwise we would have i misreport from $\succ_6$ to $\succ_7$ and thus go from being unmatched to being matched in a valid outcome. However the latter would also not be strategy-proof, as k in $\succ_2$ would misreport to $\succ_7$ and go from being unmatched to being matched in a valid outcome. Thus there is no way of choosing a valid match. Hence there is no desirable mechanism. □

A.2 Proof of Proposition 2

Note that we adopt some of the notation in Gilon et al. (2019) for a more direct comparison.

Proof. Observe the following. First, no two P_i have type AB. If this were not the case, for example $P_1 = P_2 = AB$, then every donor of P_1 and P_2 could donate to the other patient. This contradicts with the set of feasible pairwise matches, which does not have this property for any pair of patient-donor groups. This implies that $d_i^1 \neq AB$ for all i. To see this, note that for each d_i^1, they can donate to two other patients. If for contradiction we had that for some i, $d_i^1 = AB$, then we would have that there are two AB patients because AB donors can only donate to AB patients. Furthermore, we have that for $i \in \{2, 3, 4\}$, $P_i \neq AB$. To see this, consider $P_2 = AB$ for contradiction. We have that $d_2^1 - d_4^2$ is a feasible pairwise exchange. As such, d_2^1 can donate to P_4, and since all donors can donate to $P_2 = AB$, then so can d_4^1. However $d_2^1 - d_4^1$ is not a feasible pairwise exchange, which is a contradiction. We can apply the same argument to the case where $P_3 = AB$ or $P_4 = AB$ (but note that it does not apply to P_1). Finally, observe that $P_i \neq O$ for $i \in \{2, 3, 4\}$. If this were not true, for example $P_2 = O$, then we would have that $d_1^1 = d_4^2 = O$. Given that O donors can donate to any agent, we would have that $d_1^1 - d_4^2$ would be a feasible exchange. This is a contradiction with the set of feasible exchanges given. The same argument applies to $P_3 = O$ or $P_4 = O$.

Using these facts, we proceed by considering various possible cases. First consider the case that $P_1 = AB$, thus any donor can donate to P_1. This implies $d_1 \neq O$, which follows because if $d_1 = O$, then we would have that any exchange is feasible for P_1 as P_1 can receive from any donor as they are AB and d_1^1 can donate to any patient as they are O. Recall that $d_1^1 \neq AB$, hence $d_1^1 \in \{A, B\}$. Consider the case that $d_1^1 = A$, and thus $P_2 = P_3 = P_4 = A$ (because none of these patients can be AB). Hence we must have that $d_2^2 \in \{A, O\}$ because $d_2^2 - d_3^1$ is feasible and $P_3 = A$, which means $d_1^1 - d_2^2$ is feasible and thus a contradiction. A similar proof applies when we instead assume that $d_1 = B$. As we have considered all the possible cases for choice of blood type of d_1 given $P_1 = AB$, and they all lead to contradictions, we cannot have that $P_1 = AB$.

Now consider the case where $P_1 = O$. As O patients can only receive from O donors, then we have that $d_2^1 = d_3^1 = d_4^1 = O$. This would imply that $d_3^1 - d_4^1$ is feasible, which is a contradiction. Hence we must have that $P_1 \in \{A, B\}$. Recall that we also have $P_2 \in \{A, B\}$.

Consider $P_1 = P_2 = A$. Thus we must have that $d_1^1, d_4^2 \in \{O, A\}$. If $d_1^1 = O$, then this would imply that $d_1^1 - d_4^2$ is feasible, which is a contradiction. If $d_1^1 = A$, then $P_4 = A$ (as it cannot be AB). Furthermore, if d_4^2 can donate to $P_2 = A$, then it can also donate to $P_1 = A$. And since d_1^1 can donate to P_4, we have that $d_1^1 - d_4^2$ is feasible, which is a contradiction. The same idea applies to $P_1 = P_2 = B$.

Now consider $P_1 = A$ and $P_2 = B$ (the same idea applies for $P_1 = B$ and $P_2 = A$). Note that $d_1^1, d_4^2 \in \{B, O\}$, because both donors can donate to $P_2 = B$. If $d_1^1 = d_4^2 = O$, then we would have $d_1^1 - d_4^2$ is feasible, which is a contradiction. If $d_1^1 = B$ and $d_4^2 = O$, then we have that $d_1^1 - d_4^2$ is feasible, which is also a contradiction. Consider $d_1^1 = O$ and $d_4^2 = B$. Because $d_1^1 - d_4^1$ is feasible and $P_1 = A$, it must be that $d_4^1 \in \{O, A\}$. If $d_4^1 = O$, then $d_4^1 - d_2^1$ would be feasible, which a contradiction. Hence it must be that $d_4^1 = A$. Given that $d_1^1 = O$ can donate to P_2, but d_2^2 cannot donate to $P_1 = A$ because $d_1^1 - d_2^2$ is not feasible, it must be that $d_2^2 \in \{B, AB\}$. First consider $d_2^2 = B$. In this case, note that $P_4 \in \{O, A\}$ because d_4^2 is compatible with P_2 but $d_2^2 = B$ is not compatible with P_4 as $d_2^2 - d_4^2$ is not feasible. If $P_4 = O$, then we must have that $d_2^1 = O$ as $d_2^1 - d_4^2$ is feasible. But this is a contradiction because we would have that $d_3^1 - d_2^1$ is feasible. If $P_4 = A$, then this would imply that $d_2^1 = \{A, O\}$. It must be that $d_2^1 = A$, as if $d_2^1 = O$ then we would have that $d_2^1 - d_3^1$ is feasible, which is a contradiction. Given this, we must have that $P_3 \in \{B, O\}$ so that $d_2^1 - d_3^1$ is not feasible (as $d_2^1 = A$). By our previous observation, it cannot be that $P_3 = O$. Hence if $P_3 = B$, we would get a contradiction as $d_3^2 - d_4^2$ would be feasible because d_3^2 is compatible with P_4 given the set of feasible exchanges, and $d_4^2 = B$ and $P_3 = B$ by assumption. Thus we cannot have that $d_2^2 = B$, so it must be that $d_2^2 = AB$. This would imply that $P_3 = AB$, which is a contradiction with out observation that $P_2, P_3, P_4 \neq AB$.

Now we consider the final case, that is $d_1^1 = d_4^2 = B$. This implies that $P_4 \in \{B, O\}$ as $d_1^1 - d_4^1$ is feasible and $d_1^1 = B$ by assumption. If $P_4 = B$, then $d_3^2 \in \{B, O\}$ as $d_2^3 - d_4^1$ is feasible and these are the only feasible blood types that can donate to $P_4 = B$. In either case, we would have that $d_2^2 - d_2^3$ is feasible because $P_2 = B$, which is a contradiction. Thus it must be that $P_4 = O$. However this contradicts our earlier observation that $P_4 \neq O$. As such it cannot be that $P_1 = A$ and $P_2 = B$ (or by an analogous argument that $P_1 = B$

and $P_2 = A$). As we have gone through all cases, and shown that there is no choice of blood types that are consistent with this set of feasible exchanges, then this set is not rationalizable. □

A.3 Proof of Theorem 1

Individual rationality is clear. We focus on showing efficiency and strategy-proofness.

A.3.1 Strategy-proof

First we can observe that $A - O*$, $B - O*$, $b - b*$ for $b \in \{O, A, B\}$, and $B - A*$ all get their top choice and have no incentive to deviate. Furthermore, all $A - *$ agents are guaranteed their top two IR choices. Hence the only agents that do not get their top donor are $A - B, A*$ matched via A; $A - B, O*$ matched via O; and $A - B$ unmatched. If any other donor is put ahead of B, then A will be guaranteed to be matched by them. Hence there is no incentive to deviate as they will not be able to improve to their top donor instead of their second donor. Finally, $O - *$ can be seen to have no incentive to deviate by the Serial Dictator procedure, which uses a fixed priority order. Furthermore, all $O - *$ agents take the best option available to them when it is their turn.

A.3.2 Efficiency

We will construct a sequence, $i_0 \to i_1 \to i_2 \to \ldots$, where $i_n \to i_{n+1}$ means that i_n is matched with i_{n+1}'s partner in M', and i_{n+1} has a new partner in M'. We see this as i_n *stealing* i_{n+1}'s partner in M and displacing them. For three cycles, we specify a certain agent as being stolen, and a certain agent who was stolen from. We refer to the other agents as being *free*, and we should that free agents are never stolen. Note that i_0 steals first, in the sense that the agent they are pointing to can no longer use that agent. Hence this kicks off a chain of stealing from different agents where no two agents steal from the same agent. Thus there are infinitely many agents (a contradiction).

We assume without loss that in M', the following cycles don't happen: $A - B \leftrightarrow B - O$, $B - A \leftrightarrow A - O$, or any three cycle with $b - b$. We justify this as follows. For $A - B \leftrightarrow B - O$, since there is a surplus of $O - A$ we can Pareto improve M' to M'' that implements $O - A \to A - B \to B - O$. Similarly for $B - A \leftrightarrow A - O$, which can be improved to $O - B \to B - A \to A - O$. Consider a three cycle with $b - b$ of the form $X_1 - X_2 \to b - b \to Y_1 - Y_2$. Due to transitivity of the blood type compatibility relation,

X_2 can donate to Y_1. Thus we can match $X_1 - X_2 \leftrightarrow Y_1 - Y_2$ and match $b - b$ with other $b - b$ in a two or three cycle as we assume there are at least 2 such agents.

These agents are the ones that can be improved in some match M:

- $A - B, (A)*$; $A - B, (O)*$; $A - B$
- $O - A, (B)*$; $O - A, (O)*$; $O - A, B, (O)$
- $O - B, (A)*$; $O - B, (O)*$; $O - B, A, (O)$
- $O - *$

where (b) means that the b donor was used in M. if there is no (b), then they were unmatched. If there is no $*$, then they have no more donors other than those listed. If there is a $*$, then they may have more donors. $[b]$ refers to the donor used in M'.

Consider the following cases:

- $i_0 = A - [B], (A)*$. Note that they must have been matched in M via $A - A$ hence we can let $A - A$ be a free agent.
- $i_0 = A - [B], (O)*$. Note that they must have been matched in M via $O - A \leftrightarrow A - O$, $O - B, A \leftrightarrow A - O$, hence we can let $O - A$ or $O - B, A$ be a free agent.
- $i_0 = A - [(B)]$
- $i_0 = O - [B]*, O - A, [B]*, O - [A]*, O - B, [A]*$ ($O - *$ was unmatched and improved to B or A)
- $i_0 = O - [A], (O); O - [B], (O); O - [A], B, (O); O - [B], A, (O); O - A, [B], (O); O - B, [A], (O)$ ($O - *$ was matched with $O - O$ and improved to A or B, so there is an $O - *O*$ free agent)

1. If $i_n = A - [B]*, O - [B]*, O - A, [B]*$ steals
   (a) $B - (A)*$, then $i_{n+1} = A - [B]*$ (no free agents)
   (b) $B - (O)*$, then $i_{n+1} = A - [B]*, O - [B]*, O - A, [B]*, O - [A], (B)*$ (with $O - (A)*$ potentially a free agent)
2. If $i_n = O - [A]*, O - B, [A]*$ steal
   (a) $A - (O)*$, then $i_{n+1} = O - [(A)]*, O - B, [(A)]*, O - [B], (A)*$ (no free agents)

(b) $B-(O)*$, then $i_{n+1} = A-[B]*, O-[B]*, O-A, [B]*, O-[A], (B)*$ (with $O-(A)*$ potentially a free agent)

3. No free agent or unmatched (in M) agent is ever stolen, and all i_{n+1} cannot be matched with i_n in M'.

Now consider the following case:

- $i_0 = O-[A], (B)*$ (so $B-(O)*$ is a free agent). Note that there are no $O-B, (A)*$ in such a case (that is $O-B, A \leftrightarrow A-O$), as by SD procedure we must have had the $A-(O)*$ run out as $O-B$ prioritizes $B-(O)*$. Furthermore, if $O-A, B*$ had to directly match with $B-(O)*$ then there must not be an $A-B*$ unmatched as $O-A*$ prioritizes $O-A* \to A-B* \to B-O*$.

1. If $i_n = O-[A]*, O-B, [A]*$ steal
   (a) $A-(O)*$, then $i_{n+1} = O-[(A)]*, O-B, [(A)]*$ (no free agents)
   (b) $A-(B)*$, then $i_{n+1} = O-[A]*$ (and $B-(O)*$ is a free agent)
   (c) There are no $A-B$ unmatched by the previous observation.
2. No free agent or unmatched (in M) agent is ever stolen, no type of agent stolen is ever unmatched in the first place, and all i_{n+1} cannot be matched with by i_n in M'

Now consider the following case:

- $i_0 = O-[B], (A)*$ (so $A-(O)*$ is a free agent). Note that there are no $O-A, (B)*$, as either they are before i_0 in SD and would have taken $A-(O)*$, or were after serial dictator in which case i_0 would have taken $B-(O)*$.

1. If $i_n = O-[B]*$ steal
   (a) $B-(O)*$, so $i_{n+1} = O-[(B)]*, A-[(B)]*$ ($O-(A)*$ free agent). We use the fact that there are no $O-A, (B)$ here.
   (b) $B-(A)*$, so $i_{n+1} = A-[B]*$ (no free agent)
2. If $i_n = A-[B]*$ steal
   (a) $B-(A)*$, so $i_{n+1} = A-[(B)]*$ (no free agent)

(b) $B-(O)*$, so $i_{n+1} = O - [(B)]*, A - [(B)]*$ ($O-(A)*$ free agent). We use the fact that there are no $O - A, (B)$ here.

3. No free agent or unmatched (in M) agent is ever stolen, no type of agent stolen is ever unmatched in the first place, and all i_{n+1} cannot be matched with i_n in M'

A.4 Proof of Proposition 4

Proof. First consider the case where there is an AB patient. If the AB patient has a donor compatible with some other patient, then they must be pointing to someone. Since anyone can donate to AB patients, then everyone is pointing to AB. Hence there is a two-cycle. If the AB patient has no donor that can feasibly donate to another patient, then they are not part of any cycle.

Now consider the case where there is no AB patient. This means that any cycle cannot utilize AB donors, as AB donors can only donate to AB. Hence assume that there are no AB donors. Assume there is no two cycle but there is a cycle of length n.

First we will show that there cannot be an O donor as an patient's top choice, nor an O patient, for any agent in a cycle. Assume for contradiction that there is an agent i_1 that points with an O donor in the cycle. Hence they point to every agent. Since $i_n \to i_1$, hence $i_1 \leftrightarrow i_n$. Since there are no O donors, and only O donors can donate to O patients, then there can be no O patients.

We proceed by assuming that there are only A and B donors and patients, and consider different cases based on the number of agents n in the cycle.

If there are only two agents, and thus all these agents are in the cycle, then this is a contradiction.

If there are at least five agents in the cycle, denoted $i_1 \to i_2 \to i_3 \to ...$, then observe we can't have the any three consecutive agents have the same type. For example, if $i_1, i_2, i_3 \in \mathcal{I}_A$, then $i_2 \to i_1$ and thus there is a two cycle, which is a contradiction.

First consider the case where $i_1, i_2 \in \mathcal{I}_A$. This implies that $i_3 \in \mathcal{I}_B$. If the cycle is of length three, then we are done. If $i_3 \to i_4 \in \mathcal{I}_A$, then $i_3 \to i_2$, which introduces a two cycle. Hence $i_4 \in \mathcal{I}_B$. If the cycle is of length four, then we are done. If $i_4 \to i_5 \in \mathcal{I}_B$, then $i_4 \to i_3 \in \mathcal{I}_B$, another contradiction. If $i_4 \to i_5 \in \mathcal{I}_A$, then $i_4 \to i_2 \in \mathcal{I}_A$ and $i_2 \to i_4 \in \mathcal{I}_B$ since $i_2 \to i_3 \in \mathcal{I}_B$. This final contradiction shows that i_1, i_2 cannot both be in $\mathcal{I}_A$. A similar argument holds for $i_1, i_2 \in \mathcal{I}_B$.

Now consider the case where $i_1 \in \mathcal{I}_A$ and $i_2 \in \mathcal{I}_B$. If there are only three or four agents, then we are done (assuming in the latter case there is no three consecutive agents of the same type). Assume there are at least five agents: $i_1 \to i_2 \to i_3 \to i_4 \to i_5$. If $i_3 \in \mathcal{I}_A$ and $i_4 \in \mathcal{I}_B$, then $i_3 \to i_2$, giving a contradiction. By the previous argument on consecutive types, it cannot be that $i_3, i_4 \in \mathcal{I}_B$. If $i_3, i_4 \in \mathcal{I}_A$, then $i_5 \in \mathcal{I}_B$, otherwise there will be three agents of consecutive types. Then $i_4 \to i_2$ and $i_2 \to i_4$ gives a two cycle and thus a contradiction. Now consider $i_3 \in \mathcal{I}_B$ and $i_4 \in \mathcal{I}_A$, then $i_1 \to i_3$ and $i_3 \to i_1$, another contradiction. A similar argument applies for $i_1 \in \mathcal{I}_B$ and $i_2 \in \mathcal{I}_A$.

If there are exactly three agents, it is clear that the cycles must be of the following form: $A \to B \to A$, or $B \to A \to B$. Note that, for example, the latter is equivalent to $A \to B \to B$. However then it must be that there is a two-cycle given by $A \leftrightarrow B$ in both cases, thus this is not possible. If there are four agents, since there can be no three consective agents, it must be that cycles are either of the form: $A \to B \to A \to B$, or $A \to B \to B \to A$. However neither are possible as there is a two cycle, which is a contradiction.

$\square$