# The Roommates Problem Revisited

Thayer Morrill*
University of Maryland
November 2007

**Job Market Paper**

## Abstract

One of the oldest but least understood matching problems is Gale and Shapley's (1962) "roommates problem": is there a stable way to assign $2N$ students into $N$ roommate pairs? Unlike the classic marriage problem or college admissions problem, there need not exist a stable solution to the roommates problem. However, the traditional notion of stability ignores the key physical constraint that roommates require a room, and it is therefore too restrictive. Recognition of the scarcity of rooms motivates replacing stability with Pareto optimality as the relevant solution concept. This paper proves that a Pareto optimal assignment always exists in the roommates problem, and it provides an efficient algorithm for finding a Pareto improvement starting from any *status quo*. In this way, the paper reframes a classic matching problem, which previously had no general solution, to become both solvable and economically more meaningful.

*Email address: thayermorrill@gmail.com. I would like to thank Larry Ausubel, Peter Cramton, Melinda Sandler Morrill, and Daniel Vincent for their comments throughout this project. I would also like to thank Daniel Aromi and Jonah Gelbach for their helpful suggestions.

# 1 Introduction

Economics is often defined as the study of how to efficiently allocate scarce resources. As such, assignment problems are at the heart of economics. Two-sided matching theory asks how to best match agents of two distinct types. Examples include students and schools, residents and hospitals, or kidneys and people in need of a transplant. A different but related question asks how to best pair two agents of the same type. Examples of these one-sided matches include roommates at a university, lab partners in a science class, and partners in a police force. Two-sided matching theory has been well studied by economists who have created an elegant and applicable theory. One-sided matching theory has been comparatively neglected[1].

This neglect is likely due to the very paper that introduced it. In their classic 1962 article *College Admissions and the Stability of Marriage*, Gale and Shapley introduce both the marriage problem and the roommates problem. While Gale and Shapley prove a stable match always exists in a two-sided market, they introduce the roommates problem to demonstrate that a stable pairing need not exist in a one-sided market. Since a stable match need not exist, economists have been stymied in their attempts to find and analyze solutions to this important assignment problem. Unfortunately, this has led many economists to turn their attention elsewhere, and as a result, the economics literature on this classic problem is sparse.

This paper starts by questioning if stability is the correct equilibrium concept. Gale and Shapley define a set of marriages as unstable if either there exist a man and woman who are not married but prefer each other to their current spouse or there exists someone who would prefer to be single than married

[1] Roth and Sotomayor (1990) is an excellent introduction to the two-sided matching literature. Gusfield and Irving (1989) is also a nice introduction. Interestingly, although the economics literature on the roommates problem is very small, there is a comparatively large computer science literature on it. Roth and Sotomayor, two economists, mention the roommates problem only as an example. In contrast, Gusfield and Irving, two computer scientists, devote nearly a quarter of the book to the roommates problem. Finding a traditionally-stable roommate pairing (if one exists) is considered a "hard" algorithmic question. The bulk of their presentation is a polynomial-time algorithm for finding a traditionally-stable pairing when one exists. Tan (1991) establishes a necessary and sufficient condition for the existence of a stable pairing. Chung (2000) extends Tan's result to a sufficient condition for the existence of a stable pairing when preferences are weak.

to their current partner. Stability in the roommates problem is borrowed from the marriage model. A pairing is unstable if two students prefer to live with each other rather than their current assignment.[2] Stability fits the marriage model so well that no other solution concept has been needed or suggested. The same is not true of the roommates problem. Roommates face an additional constraint that married couples do not; roommates must have a room in which to live. A student may prefer another to her assigned roommate; however, she needs a room in which to live and presumably does not have the right to evict her current roommate. Therefore, the traditional notion of stability is too restrictive.

I will present Gale and Shapley's original example to highlight this point.

**Example (Gale and Shapley, 1962):** *A Stable Assignment Need Not Exist*

*Suppose there are four students: $\alpha, \beta, \gamma$ and $\delta$. $\alpha$'s top choice is $\beta$, $\beta$'s top choice is $\gamma$, $\gamma$'s top choice is $\alpha$, and all three rank $\delta$ last. Gale and Shapley define an assignment to be unstable if two students are not currently roommates but prefer each other to their current assignment. Under this definition, there does not exist a stable assignment since whoever is assigned to $\delta$ prefers the other two students to $\delta$ and is the top choice of one of these students. In the words of Gale and Shapley:*

> "...whoever has to room with $\delta$ will want to move out, and one of the other two will be willing to take him in."

While one of the other two may be willing to take him in, it is quite a different matter whether this student is *able* to take him in. In order to take him in, either his current roommate must voluntarily leave, be evicted, or an additional room must be available. With a scarcity of rooms and with no student willing to change his assignment to $\delta$, the original assignment is an equilibrium after all.

[2] I am interested in the case where each student is required to have a roommate. Consequently, I do not include in my definition of stability the additional requirement that each student prefers her assignment to being unassigned.

If an agent can dissolve her partnership unilaterally, then stability is the natural equilibrium concept. If she finds someone she prefers who also prefers her, then both parties will dissolve their current partnership and pair together. As a result, the original assignment is not an equilibrium. However, if a partnership requires *bilateral* agreement to dissolve, then two agents wanting to change their assignment is not enough to disturb the original pairing. If bilateral agreement is required, an assignment will only be changed if all involved parties agree. Since an agent will only agree if the new assignment makes her better off, any deviation from the original set of assignments must be a *Pareto improvement*. Therefore, when bilateral agreement is required to dissolve a partnership, Pareto optimality, not stability, is the proper equilibrium concept. If an assignment is Pareto optimal, then there is no reassignment that all parties will consent to; therefore, the original assignment will not be disturbed.

Most of matching theory studies assignments that can be unilaterally dissolved. Assignments which can only be dissolved with bilateral agreement are an important but little studied second category. As argued above for the roommates problem, an essential but scarce input creates the need for bilateral agreement. Additional examples include police officers who require a police car to do patrol and partners in a science class who must work at a common laboratory. The same requirement can be created by a legally binding contract that can only be modified by mutual consent. For example, many professional athletes have no-trade clauses in their contract which they may waive at their discretion. In the presence of such clauses, the assignment of an athlete to a team can only be disturbed when all relevant parties approve the trade.

This paper focuses on the roommates problem as reconsidered using the equilibrium concept of Pareto optimality. I will show there always exists an efficient assignment. Therefore, unlike the case where stability is applied, an equilibrium always exists in the roommates problem. Moreover, I show an inefficient assignment can always be Pareto improved to an efficient one. These results motivate several questions. If an assignment has not been made, how should we make it? If an assignment has been made, how can we determine if the assignment is efficient? If an assignment is inefficient, how can we Pareto improve it? These questions are the focus of this paper. In particular, the last two turn out to be complicated. To answer them I introduce an algorithm, The Roommate Swap, which identifies whether an assignment is inefficient and finds a Pareto improvement when it is.

Much of the analysis in this paper relies on tools from graph theory. Networks are a natural way of representing assignment problems, particularly one like the roommates problem where two agents are paired. In particular, my algorithm relies heavily on applying Edmund's Blossom algorithm[3] to the graph theoretic representation of the roommates problem.

The paper is organized as follows. Section 2 formally introduces the problem and proves existence. Section 3 details the Roommate Swap algorithm. Section 4 examines the strategic implications of several assignment mechanisms. Section 5 looks at extensions and modeling issues, and section 6 concludes. The appendix provides several technical proofs and a discussion of the computational complexity of the Roommate Swap algorithm.

## 2 The Roommates Problem Revisited

We wish to assign $2N$ students to $M$ rooms. Students have preferences over all other students that are strict, complete, and transitive. All rooms are identical and students have no preference as to which room they are assigned.

An assignment is a function that pairs students. Every student is assigned to exactly one other student, and assignments are symmetric.

**Definition 1.** *Let $S$ be a set of students with $|S| = 2N$. A function $\mu : S \to S$ is an* ***assignment*** *of $S$ if:*

1. $\mu(s) \neq s$.
2. $\mu(s_1) = \mu(s_2) \Rightarrow s_1 = s_2$.
3. $\mu(\mu(s)) = s$.

The traditional equilibrium concept is based on the notion of a blocking pair.

[3] Edmunds (1965).

**Definition 2.** *Two students $s$ and $t$ are a* ***blocking pair*** *to an assignment $\mu$ if $\mu(s) \neq t$ but $s \succ_t \mu(t)$ and $t \succ_s \mu(s)$. An assignment is* ***stable*** *if there does not exist a blocking pair*[4]*.*

As argued in the introduction, this is not the proper equilibrium concept for the roommates problem. A roommate assignment is an equilibrium if it is Pareto optimal.

**Definition 3.** *An assignment $\mu$ is* ***inefficient*** *if there exists a different assignment $\mu'$ such that for every student $s$, $\mu'(s) \succeq_s \mu(s)$. An assignment is* ***Pareto optimal (efficient)*** *if it is not inefficient.*

Since preferences are strict, if $\mu'$ Pareto improves $\mu$, then at least four students must strictly prefer $\mu'$ to $\mu$. As the following result proves, the set of stable assignments is a subset of the set of efficient assignments.

**Proposition 1.** *If an assignment is stable, then it is Pareto efficient.*

*Proof.* I will prove the contrapositive. If an assignment $\mu$ is inefficient, then there exists an assignment $\mu'$ that Pareto improves $\mu$. Let $s$ be any student such that $\mu(s) \neq \mu'(s)$. Since $\mu'$ is a Pareto improvement of $\mu$, both $\mu'(s) \succ_s \mu(s)$ and $s \succ_{\mu'(s)} \mu(\mu'(s))$. Therefore, $s$ and $\mu'(s)$ form a blocking pair to $\mu$. □

Note that the reverse direction need not hold. An assignment can be Pareto efficient but not be stable. In Gale and Shapley's original example, no assignment is stable yet every assignment is Pareto efficient.

With the following assumptions, the general case of 2N students and M rooms reduces to the more familiar case of 2N students and N rooms:

**Assumption 1.** *Each student prefers having a room to not having a room.*

**Assumption 2.** *Each student would rather have a room to herself than to be assigned a roommate.*

[4] The traditional definition of stability also includes the constraint that the person prefers her assignment to being unassigned. In my model every student must be assigned to some room, so I omit this additional constraint.

**Assumption 3.** *At most two students can be assigned to a room.*

Note that if $N > M$, some students will not be assigned a room. Such a student cannot be involved in a Pareto improving switch by Assumption 1. Similarly, if $N < M$, a number of students will not be assigned a roommate. Assumption 2 implies such a student will never be involved in a Pareto improving switch. Therefore, the only set of students relevant for this problem are those who have been assigned a roommate. By Assumption 3, this set has exactly twice as many students as rooms. Without loss of generality, for the rest of the paper I will assume there are $2N$ students and $N$ rooms.

Gale and Shapley show that an assignment without a blocking pair need not exist. However, an efficient assignment always exists.

**Proposition 2.** *An efficient roommate assignment always exists.*

*Proof.* (Random serial dictatorship[5]) Assign every student a priority (randomly or otherwise). Assign the student with highest priority her most preferred roommate and remove them both from consideration. From students who remain, assign the student with highest priority her most preferred roommate among those students that are unassigned. Remove these two from consideration and repeat until no students remain. This assignment is Pareto efficient. To see this, note that if a student is involved in a Pareto improvement, then necessarily her roommate must be involved as well. The student with highest priority, $s_1$, receives her top choice, $s_2$, so neither she nor her choice can be involved in a Pareto improvement. Let $s_3$ be the student who chooses second. Since neither $s_1$ or $s_2$ are involved in any Pareto improvements, if $s_3$ is part of a Pareto improvement she must be reassigned to a student among $S \setminus \{s_1, s_2\}$. However, $s_3$ already receives her top choice among this set. Therefore, $s_3$ (and consequently the student she chooses) is not part of any Pareto improvement. Similarly, the student who chooses third is not part of any Pareto improvement, and so on. □

[5] Abdulkadiroglu and Sonmez (1998) is a very nice paper on the Random Serial Dictatorship mechanism. They analyze it in the context of a housing allocation problem where $n$ students are to be assigned to $n$ rooms, but it is rather interesting how robust the Random Serial Dictatorship is. The same mechanism can be used to make a Pareto efficient assignment of a student and a room, two students to be roommates, three or more students to be roommates, students to be roommates and the room they will live in, etc.

The following is a stronger statement and implies Proposition 1. It is stated to motivate the Roommate Swap algorithm.

**Proposition 3.** *If an assignment $\mu$ is inefficient, there exists an efficient assignment $\mu'$ which Pareto improves $\mu$.*

The proof is straightforward but is included as it motivates the need for the Roommate Swap Algorithm.

*Proof.* Let $\mu$ be an assignment and $PI(\mu)$ be the set of strict Pareto improvements of $\mu$. Transitivity of preference implies $\forall \mu' \in PI(\mu), PI(\mu') \subseteq PI(\mu)$. Since $\mu' \in PI(\mu) \setminus PI(\mu'), PI(\mu') \subset PI(\mu)$. Since there are only a finite number of possible assignments, the following chain must converge to the empty set:

$$PI(\mu) \supset PI(\mu_1) \supset PI(\mu_2) \supset \ldots, \text{ where } \mu_i \in PI(\mu_{i-1})$$

In particular, there must exist an $j$ such that $PI(\mu_j) = \emptyset$. $\mu_j$ is an efficient assignment which Pareto improves $\mu$. □

Put simply, if $\mu$ is not efficient, there exists a Pareto improvement $\mu_1$. $\mu_1$ is either efficient or can be Pareto improved to $\mu_2$, etc. We must eventually reach an efficient assignment, and since preferences are transitive, this assignment must Pareto improve $\mu$.

Propositions 2 and 3 motivate two distinct but related problems. The first problem is how to make an efficient assignment when no assignment has yet been made. The second is how to Pareto improve an inefficient assignment to an efficient one. Although these two problems are very similar, it is surprising how different these processes end up being. The serial dictatorship used in Proposition 2 to show existence provides a linear-time procedure for finding an efficient assignment. In contrast to the ease of finding an efficient assignment, it is rather difficult to even determine if any given assignment is efficient let alone how to improve it. Preferences between students need not interact when assigning students, but they interact directly when determining if one assignment Pareto improves another. This makes it significantly more complicated to determine if an assignment is efficient than it is to simply find an efficient assignment.

| Students | Number of Possible Assignments |
|---|---:|
| 2 | 1 |
| 4 | 3 |
| 6 | 15 |
| 8 | 105 |
| 10 | 945 |
| 12 | 10,395 |
| 14 | 135,135 |
| 16 | 2,027,025 |
| 18 | 34,459,425 |
| 20 | 654,729,075 |
| 30 | 6,190,283,353,629,370 |
| 2N | $\frac{(2N)!}{2^N(N!)}$ |

At this point the reader may object as there is an obvious and trivial algorithm to determine if an assignment is efficient. Namely, one could simply look at each possible reassignment and determine if it Pareto improves the original. If no assignment Pareto improves the original, then the original is efficient. Unfortunately, this algorithm is of no practical use as the growth of the number of assignments relative to students being assigned is factorial. Specifically, given 2N students there exists $\frac{(2N)!}{2^N(N!)} = (2N-1)(2N-3)(2N-5)\cdots(3)(1)$ many ways of assigning them to be roommates.[6] Even for small N, this is prohibitively large. For example, there exists on order of 6 quadrillion ($6 \times 10^{15}$) many ways to assign 30 students to be roommates. Therefore, a more sophisticated process is required.

# 3 The Roommate Swap Algorithm

This section demonstrates an $O(n^2)$ algorithm for determining if an assignment is efficient. Moreover, when an assignment is inefficient I provide an $O(n^3)$ algorithm, The Roommate Swap, for finding a Pareto improvement.[7]

[6] A short proof appears in the Appendix.

[7] A discussion on the computational complexity of the algorithm appears in the appendix.

Much of the analysis uses tools from graph theory, so it is necessary to present some definitions and results. This document is intended to be self-contained, but I refer the reader to *Introduction to Graph Theory*, second edition, by Douglas West for a more detailed analysis of graph theory.

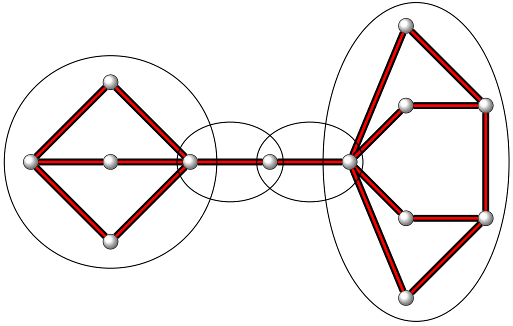
A graph consists of vertices and edges between them. For my purposes, all edges are undirected.

1. Two vertices are **adjacent** if there is an edge between them.
2. The **degree** of a vertex v, denoted $d(v)$, is the number of vertices it is adjacent to.
3. A **path** is a sequence of vertices $\{v_1, v_2, \ldots, v_k\}$ such that no vertex appears twice and any two consecutive vertices are adjacent.
4. A **cycle** is a sequence of vertices $\{v_1, v_2, \ldots, v_k\}$ such that no vertex appears twice, any two consecutive vertices are adjacent, and $v_1$ and $v_k$ are adjacent.
5. Two vertices are **connected** if there is a path between them. Since our graphs are undirected, connected is a reciprocal relationship. A graph is connected if all vertices are connected.
6. A vertex is **incident** to an edge if it is one of the edge's endpoints. $G \setminus v$ is the graph that results from deleting the vertex v and all edges incident to v.
7. A vertex v is a **cut-vertex** if G is connected, but $G \setminus v$ is not.
8. A **block** is a maximal subgraph containing no cut vertex.

Note that the subgraph consisting of two vertices and an edge between them contains no cut-vertex, so any edge is either a block or a subset of a block. I will refer to any block containing only two vertices as a trivial block. Since every vertex in our graph has at least one edge, this is the smallest block possible. Figure 1 shows an example where the blocks have been circled.

**Definition 4.** *Given an assignment $\mu$, a set of students $X$ is* ***closed under roommates*** *if $s \in X$ implies $\mu(s) \in X$.*

Figure 1: An example of a graph with four blocks.



Given a set of preferences $\succ$ and assignment $\mu$, I will induce a graph, $G^{\mu}_{\succ}$, as follows:

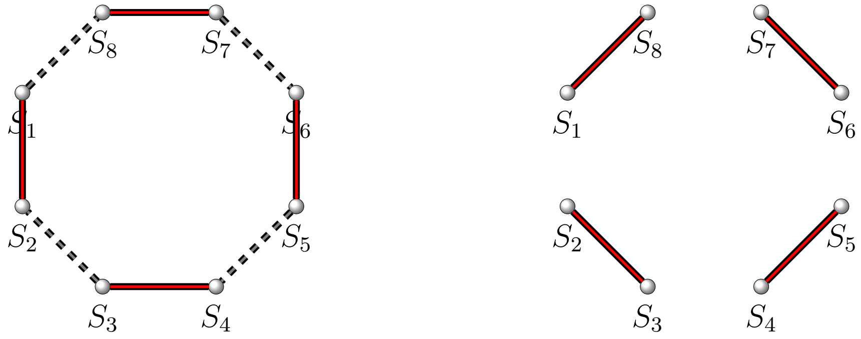
- Each vertex corresponds to a student. Label the vertices $s_1$ through $s_n$. When referring to the graph, I will use the term vertex and student interchangeably.
- A *solid* edge is drawn between roommates. By definition, each vertex is incident to exactly one solid edge.
- Draw a *dashed* edge between any two students that form a blocking pair to $\mu$. That is to say, if $s_i$ prefers $s_j$ to her current roommate and vice versa.

When the preferences and assignment are clear from the context, I will just refer to the graph as G. I will call a path that alternates between dashed and solid edges (or vice versa) an **alternating path**. Similarly, a cycle that alternates between dashed and solid edges is an **alternating cycle**.

**Lemma 1.** *An assignment $\mu$ is efficient under preferences $\succ$ if and only if $G^{\mu}_{\succ}$ contains no alternating cycle. Moreover, if $\mu'$ Pareto improves $\mu$ and $s$ is a student such that $\mu(s) \neq \mu'(s)$, then $s$ is contained in an alternating cycle in $G^{\mu}_{\succ}$.*

The intuition for sufficiency is captured in Figure 2. In an alternating cycle, we can simply "swap" roommates. We eliminate the solid edges, make

Figure 2: An alternating cycle with its corresponding Pareto improvement.



the dashed edges in the cycle solid, and leave everyone outside the cycle unchanged. This is a well-defined reassignment that Pareto improves the initial assignment.

*Proof.* Suppose $G^{\mu}_{\succ}$ contains an alternating cycle C. An alternating cycle is closed under roommates as each vertex is incident to a solid edge in the cycle. This implies $V(G) \setminus C$ is closed under roommates as well (V(G) means the vertex set of G). We will construct a Pareto improvement $\mu'$. For every $v \in V(G) \setminus C$ let $\mu'(v) = \mu(v)$. This is well defined since $V(G) \setminus C$ is closed under roommates. For every $v \in C$, let $\mu'(v)$ be the vertex it shares a dashed edge with in the cycle C. This is well defined as each vertex is incident to exactly one dashed edge in the cycle and sharing a dashed edge is a reciprocal relationship. A dashed edge indicates that both vertices prefer each other to their original assignment. Therefore, $\mu'$ Pareto improves $\mu$.

Suppose that $\mu'$ is a Pareto improvement of $\mu$. Let $G'$ be the subgraph consisting of all solid edges in $G^{\mu}_{\succ}$ and only the dashed edges between vertices not paired by $\mu$ that are paired by $\mu'$ (since $\mu'$ is a Pareto improvement, there must be a dashed edge between such vertices). Note that any vertex v in $G'$ either has degree[8] 1 (if $\mu(v) = \mu'(v)$) or degree 2 (if $\mu(v) \neq \mu'(v)$). Moreover, for any vertex v, if $d(v) = 2$, then $d(\mu(v)) = d(\mu'(v)) = 2$. Choose any vertex t such that $d(t) = 2$. $t$ is connected via a solid edge to $\mu(t)$. Since $d(t) = 2$, $d(\mu(t)) = 2$ and so $\mu(t)$ must be connected via a dashed edge to $\mu'(\mu(t))$.

[8] The degree of a vertex v, denoted $d(v)$, is the number of edges v is incident to.

$\mu'(\mu(t))$ must be connected via a solid edge to $\mu(\mu'(\mu(t)))$ which must be connected to a dashed edge via $\mu'(\mu(\mu'(\mu(t))))$, and so on. Eventually this process must cycle as there is only a finite number of vertices. However, a cycle to any vertex $s \neq t$ would mean the degree of s is at least three which is not possible. Therefore, the process must cycle back to our first vertex t. Moreover, it must cycle via a dashed edge as we have already exhausted t's solid edge. By construction, this is an alternating cycle. □

**Lemma 2.** *Let $t$ be any student.*

1. *$t$ and $\mu(t)$ are contained in a unique block, $B_t$.*
2. *If $t$ is part of an alternating-cycle $C$, then $C \subseteq B_t$.*
3. *If $t$ is involved in a Pareto improvement, then $B_t$ is non-trivial. That is to say if there exists an assignment $\mu'$ such that $\mu'$ Pareto improves $\mu$ and $\mu'(t) \neq \mu(t)$, then $|B_t| > 2$.*

*Proof.*

1. Since there is an edge between $t$ and $\mu(t)$, they are in at least one block together. Since the intersection of two blocks contains at most one student,[9] $t$ and her roommate must be in exactly one block together. Call this block $B_t$.
2. A cycle contains no cut-vertex, so it must be a subset of a block. An alternating-cycle containing $t$ must contain $\mu(t)$ since $t$ lies on a solid edge in the alternating-cycle. Since $B_t$ is the unique block containing $t$ and $\mu(t)$, the cycle must be contained in $B_t$.
3. If $t$ is involved in a Pareto improvement, then by Lemma 1 $t$ is contained in an alternating-cycle. By (2) this alternating-cycle is contained in $B_t$, so $B_t$ must contain more than just $t$ and $\mu(t)$.

[9] See West pg. 156. The intuition is that if if two blocks $B_1$ and $B_2$ share two vertices, then after cutting a vertex, at least one of the two must remain. Call this vertex $v$. $v$ is connected to all remaining vertices as it is in a block with each of them. But if every vertex has a path to $v$, then all vertices are connected. Therefore $B_1 \cup B_2$ has no cut-vertex contradicting the maximality of a block.

□

Lemma 2, part (2) says that if a student $t$ is part of a Pareto improvement (and consequently an alternating-cycle), then she must be reassigned to a member of $B_t$. Therefore, no edge between $t$ and a vertex outside of $B_t$ can be part of an alternating-cycle. Let $G'$ be the graph obtained by deleting all edges between $t$ and any vertex not in $B_t$. Then $G$ contains an alternating-cycle if and only if $G'$ contains an alternating-cycle. This motivates the following procedure.

## Pruning a Graph

1. Start with a graph G.
2. Determine the set of blocks $B_1, B_2, \ldots, B_m$.
3. For each student-roommate pair $s$ and $\mu(s)$, locate the unique block that both are in. Remove *all* edges from either $s$ or $\mu(s)$ to any student outside this block.

A key point is that if a student $s$ was in a block B with $\mu(s) \notin B$, then after pruning the graph, $s$ is no longer in B. By iterating the pruning process we end up with a graph in which all blocks are closed under roommates. Note that these blocks may be trivial, but by Lemma 2, the students in such a block are not involved in any Pareto improvements.

**Proposition 4.** *Any non-trivial block closed under roommates contains an alternating cycle.*

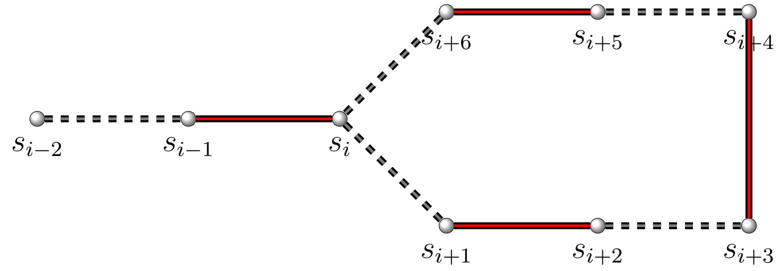
The algorithm in this proof was inspired by Edmunds' Blossom Algorithm from graph theory[10] and Gale's Top-Trading Cycles Algorithm.[11]

*Proof.* Look at any non-trivial block $B$ closed under roommates. Every vertex $v$ in $B$ must be incident to a dashed edge. Otherwise $v$ is only connected

[10] Edmunds (1965). A discussion of the Blossom algorithm appears in West, page 142.

[11] Shapley and Scarf (1974).

Figure 3: A "Blossom".



(by a solid edge) to $\mu(v)$ which would mean removing $\mu(v)$ disconnects $v$ from the rest of the block. This is not possible since a block contains no cut-vertices. Start with any vertex $s$. First take a dashed edge to a new vertex $s_1$ then continue on a solid edge to $s_2 = \mu(s_1)$. Continue alternating between dashed and solid edges until we cycle. We must eventually cycle since there is a finite number of vertices.

If our cycle is even (a cycle is even if it contains an even number of vertices), then we are done. By construction, an even cycle alternates between dashed and solid edges and is therefore an alternating cycle. Therefore, assume our cycle is odd, $\{s_i, s_{i+1}, s_{i+2}, \ldots, s_{i+2m}\}$. By construction, any odd cycle looks like Figure 3, except possibly of different length. Edmunds refers to this as a blossom. The vertices $\{s_1, s_2, \ldots, s_i\}$ are the stem, $s_i$ is the base of the blossom, and $s_i$ must connect to $s_{i+1}$ and $s_{i+2m}$ via dashed edges.

There must by a dashed edge from one of $s_{i+1}, s_{i+2}, \ldots, s_{i+2m}$ to a vertex outside the cycle. Otherwise $s_i$ would be a cut-vertex as deleting it would disconnect $s_{i+1}, s_{i+2}, \ldots, s_{i+2m}$ from the rest of the graph. What we will do is contract the odd cycle into a single super-vertex $S_i^1$. The superscript indicates the number of contractions we performed to result in $S_i$. See Figure 4 for an example. Make any edge that was previously between a vertex in the cycle and a vertex $t$ outside the cycle now between $S_i^1$ and $t$. Note that there is a solid edge between $s_{i-1}$ and $S_i^1$ and all other edges incident to $S_i^1$ must be dashed as for any $s_j \in \{s_{i+1}, s_{i+2}, \ldots, s_{i+2m}\}$, $\mu(s_j) \in \{s_{i+1}, s_{i+2}, \ldots, s_{i+2m}\}$.

Now continue with one of the unexplored dashed edges incident to $S_j^1$. Again, we must eventually cycle. If the cycle is even, stop. If the cycle is odd, contract the blossom and continue. There must always be an unexplored dashed edge out of an odd cycle (or else the base vertex would be a cut vertex), so after any odd cycle we will be able to continue. Since we have a finite number of vertices and edges and each contraction reduces the number of vertices, we cannot continue indefinitely. The algorithm only stops with an even cycle, and since the algorithm must eventually terminate, we must eventually reach an even cycle.

Any alternating cycle containing super-vertices can be expanded to an alternating cycle containing no super-vertices. No matter how we enter the blossom, either the edge to the left or to the right is solid. We can follow the cycle in the direction of the solid edge all the way to base vertex. This is an alternating path to the base, the cycle connects to the base with a dashed edge, and then continues along the stem starting with a solid edge. So indeed, this expands an alternating path through a super-vertex to an alternating path through the cycle that was contracted. If our super-vertex is the result of multiple contractions, then our base vertex is now a super-vertex but otherwise nothing changes. Moreover, the base is a super-vertex containing fewer contractions, so we can proceed by induction on the number of contractions to get the desired result.

□

Proposition 4 implies a simple procedure for determining whether an assignment is efficient.

**Determining if an assignment $\mu$ is efficient given preferences $\succ$.**

1. Induce graph $G^{\mu}_{\succ}$.

Figure 4: A blossom before and after contraction.

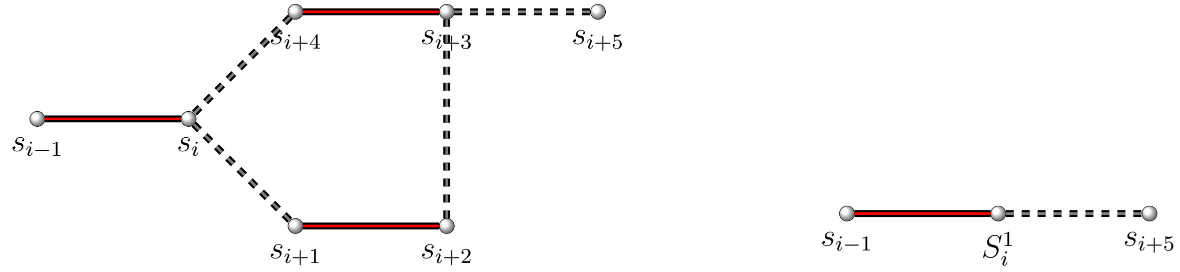
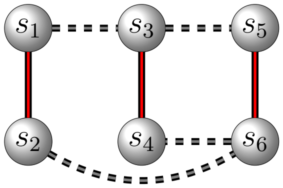


Figure 5: A non-trivial block closed under roommates, but $s_1$ and $s_2$ are not contained in any alternating-cycle.



2. Iteratively prune $G^{\mu}_{\succ}$ until all blocks are closed under roommates.

3. If all blocks are trivial, then our assignment is efficient. If there exists a non-trivial block, then by Proposition 4 and Lemma 1, our assignment is inefficient.

The algorithm in the proof finds an alternating cycle when one exists. Once we have located an alternating cycle then just as we did in Figure 2 on page 11, we "swap" roommates to get a Pareto improvement, . For this reason I call the algorithm the Roommate Swap. Note that we have now answered the two key questions from the previous section. The Roommate Swap identifies if a given assignment is efficient. Moreover, when an assignment is inefficient, it finds a Pareto improvement.

The Roommate Swap determines if a given assignment is efficient. However, a particular student likely does not care whether the assignment can be Pareto improved. Rather, she would like to know if *she* can be part of a Pareto improvement. Unfortunately, Proposition 4 does not generalize to the statement if a student $t$ is contained in a non-trivial block closed under roommates, then $t$ is involved in a Pareto improvement. Figure 5 is a non-trivial block that is closed under roommates, but $s_1$ and $s_2$ are not part of any Pareto improvements.

The Roommate Swap does not determine if a particular student can improve her assignment. However, it is not biased. If we randomly choose the vertex we start with, and when we have a choice, we randomly choose which edge to continue on, then the Roommate Swap will find any Pareto improvement with probability that is uniformly bounded away from zero. Therefore, if the

Roommate Swap is run repeatedly, it will determine if an individual student is involved in a Pareto improvement with probability one.

# 4 Strategic Implications

This paper has focused on two problems: finding an efficient assignment and finding an efficient Pareto improvement of an inefficient assignment. Continuing the pattern from previous sections, finding a strategy-proof mechanism for making an assignment is easier than finding a strategy-proof mechanism for improving an assignment. In fact, we will find that there does not exist a mechanism for selecting a Pareto improvement of a given assignment that makes truthful revelation of preferences a dominant strategy. These results follow very closely the results for two-sided matching theory presented in Roth and Sotomayor (1990).

Following the matching literature, I will use dominant strategy as my equilibrium concept.

**Definition 5.** *A **dominant strategy** is a strategy that is a best response to all possible strategies of the other agents. An assignment mechanism is **strategy proof** if it is a dominant strategy for each agent to reveal her preferences truthfully.*

There does exist a strategy-proof mechanism for making an efficient assignment. In fact, we have already seen this mechanism several times.

**Observation 1.** *The serial dictatorship is strategy proof.*

In the serial dictatorship, a student's preferences are irrelevant unless she is the one choosing her roommate. Since she gets her top choice, she does best when she submits her true preferences regardless of the preferences submitted by other students.

Finding an incentive compatible, efficient assignment mechanism is very closely related to Social Choice theory and Arrow's Impossibility Theorem. The Gibbard-Satterthwaite Theorem says that if arbitrary preferences are possible, then the unique incentive-compatible, Pareto optimal mechanism is the dictatorship mechanism. Unfortunately this cannot be directly applied as we are restricting the domain of allowable preferences. A students is only allowed to have preferences over her own assignment, and therefore, she is forced to be indifferent between many assignments. For example, a student does not have a single most-preferred assignment, but rather, she is indifferent among all assignments that match her to her most-preferred roommate. A dictator mechanism would not be Pareto optimal as, among her top choices, the dictator would select a Pareto optimal assignment only by chance. The serial dictatorship is the generalization of the dictatorship mechanism that has the properties of incentive compatibility and Pareto optimality. Due to the corresponding uniqueness results for the dictatorship mechanism, it seems likely that the serial dictatorship is the unique incentive-compatible mechanism for selecting an efficient assignment.

**Lemma 3.** *There does not exist a strategy-proof mechanism for selecting a Pareto improvement of an inefficient assignment.*

Lemma 3 is proved in the appendix. This is quite a general result, but it is rather easy to proof. A strategy-proof mechanism must be able to handle any initial assignment and any profile of preferences. Following the path of Roth (1982), I demonstrate a case that no mechanism is able to handle.

# 5 Extensions and Modeling Issues

## 5.1 Extensions to the Model

Not surprisingly, the existence of an efficient solution is quite general. For example, if students have preferences over both their roommate and the room they are assigned, then Propositions 2 and 3 still hold. In fact, the same proofs are still valid. Similarly, if more than two students are assigned to be roommates, the same existence results hold.

This paper has focused on one-sided matches, but there are many interesting examples of two-sided matches with a physical constraint. Whenever a two-sided match requires bilateral approval to dissolve, then any Pareto optimal assignment will be an equilibrium. For example, an airline matches a pilot with a navigator in order to fly an airplane. The presence of a physical constraint, the airplane, means a blocking pair is not enough to disturb an assignment.

The extra structure inherent in a two-sided match makes it easier to find a Pareto improvement to a two-sided match than a one-sided match. Here we can use a slight variation of the Top Trading Cycles algorithm[12] to determine if an assignment is Pareto optimal and to Pareto improve the assignment when it is not. For a given pilot p, define a navigator n to be *achievable for p* if n weakly prefers p to her current assignment. Have each pilot point to her most-preferred, achievable navigator. Note that a pilot always has a navigator to point to as her current assignment is achievable. Have each navigator point to their current assignment. There must exist a cycle since there are only a finite number of agents and each agent is pointing to someone. If the cycle is trivial (the pilot is pointing to the navigator she is currently assigned to), then neither the pilot nor the navigator can be involved in a Pareto improvement and we can remove them from consideration. If the cycle is non-trivial, then it represents a strict Pareto improvement for all agents in the cycle. Future drafts of this paper will contain a more detailed discussion of two-sided matches with a physical constraint.

When students have preferences over both their roommate and the room they are assigned, there still exists an efficient assignment. However, the Roommate Swap does not readily generalize to this case. The notion of a "swap" completely characterizes a Pareto improvement when only one other factor is involved in a pairing; however, with multiple dimensions a Pareto improvement can be much more complicated.

However, there is one very specific but important case where the Roommate Swap can be readily generalized. If students have lexicographical preferences over their roommate and room, then we will be able to find a Pareto improvement for any inefficient assignment. If the students care about the room first and the roommate second, then we can run the Top Trading Cycles algorithm to find a Pareto improvement when one exists. If a student cares about her roommate first and her room second, then we can first run the Roommate Swap and next run the Top Trading Cycles algorithm. There

[12] See Shapley and Scarf, 1974.

are a variety of ways we can aggregate individual preferences over rooms to a single roommate-pairing preference over rooms that will result in an efficient allocation. Thus, starting with an arbitrary assignment and lexicographical preferences over roommates and rooms, we can determine if an assignment is efficient, and if not, Pareto improve it to an efficient one.

## 5.2 Alternative Equilibrium Concepts

This paper has focused on pairing two agents as this is the classic framing of the roommates problem. While I believe efficiency, not stability, is the correct equilibrium concept for this classic problem, the more agents that are assigned to be together, the less compelling efficiency becomes as an equilibrium. If six people are assigned to an office, it is likely that a person can switch desks with a student in another office without requiring unanimous approval from her current officemates.

We can formulate an alternative equilibrium that has more appeal in this case. Instead of assigning six people to be officemates, we make six keys to each office and give each person a key to one office. Since the rooms are homogeneous, this is just an indirect way of assigning officemates. If we allow students to trade keys, then an assignment is an equilibrium if no two students wish to trade offices. Note that we are honoring the physical constraint; no student is being evicted. Moreover, this does not allow a student to block another student from switching her assignment.

Similar to stability, there need not exist an equilibrium if students can trade rooms. Suppose there are four students, a, b, c, and d, with preferences as follows:

$$\begin{array}{rcl} a & : & b \text{ is most preferred} \\ b & : & c \text{ is most preferred} \\ c & : & d \text{ is most preferred} \\ d & : & a \succ c \succ b \end{array}$$

If $a$ is assigned to $b$ and $c$ is assigned to $d$, then $b$ and $d$ will trade places. If $a$ is assigned to $d$ and $b$ is assigned to $c$, then $a$ and $c$ will trade places. Finally, if $a$ is assigned to $c$ and $b$ is assigned to $d$, then $a$ and $d$ will trade places. Since these are the only possible assignments, there is no equilibrium.

In general, an argument can be made for either equilibrium. the new equilibrium might be a reasonable model for condominiums or rooms in a group house. If a person decides to sell her condominium, she does not need the approval of the other condominium owners in the building. Note however that there also exists building cooperatives. Here a sale does require the approval of the board, so in this context, Pareto optimality is a more natural equilibrium concept. Similarly, depending on the lease a person signs, subletting a room in an apartment or group house may or may not require the approval of the landlord or other tenants. Therefore, whether or not Pareto optimality is the best equilibrium concept depends on the particular lease signed.

# 6 Conclusion

The roommates problem is one of three assignment problems introduced by Gale and Shapley in their classic 1962 paper *College Admissions and the Stability of Marriage*. This is the paper that created the field of matching theory, and the reason why the roommates problem was included is that it is such a natural assignment problem. While their other two assignment problems, the marriage problem and the college admission problem, have been studied extensively, little progress has been made on the roommates problem. This paper hopes to make several contributions to the matching theory literature on the roommates problem. First, identifying Pareto optimality instead of stability as the proper equilibrium makes the roommates problem economically more meaningful. With this improved equilibrium concept, I have shown that an equilibrium always exists. Most importantly, I demonstrate how to improve an inefficient assignment to an efficient one if we are not in equilibrium. For such a natural assignment problem as the roommates problem, this is likely to have real world applications. Therefore, this paper reframes a classic matching problem, which previously had no general solution, in a way that is both solvable and economically more meaningful.

# 7 Appendix

**Lemma 4.** *There are $\frac{(2N)!}{2^N(N!)} = (2N-1)(2N-3)(2N-5)\cdots(3)(1)$ many ways to assign 2N students to be roommates.*

*Proof.* The proof is by induction. When $N = 1$, the result is trivial as there is only one way to assign two students to be roommates. Assume $N > 1$ and by induction there are $(2N-3)(2N-5)\cdots(3)(1)$ many ways to assign 2(N-1) many students to be roommates. Select a student s. There are 2N-1 possible roommates for s, and by assumption, for any roommate we pick, there are $(2N-3)(2N-5)\cdots(3)(1)$ many ways to assign the remaining 2N-2 many students. Therefore, there is a total of $[2N-1]\times[(2N-3)(2N-5)\cdots(3)(1)]$ many ways of assigning roommates. □

**Lemma 3** *There does not exist a strategy-proof mechanism for selecting a Pareto improvement of an inefficient assignment.*

*Proof.* Suppose there are four students, a, b, c, and d, and an initial assignment, $\mu_1$ pairing a with b and c with d. Moreover, suppose the student's preferences are as follows.

$$\begin{aligned} a &: c \succ d \succ b \\ b &: c \succ d \succ a \\ c &: b \succ a \succ d \\ d &: b \succ a \succ c \end{aligned}$$

With four students, there are three possible assignment. Note that an assignment is completely determined by who a (or any other student) is assigned to. Let $\mu_2$ denote the assignment where a is paired with c and $\mu_3$ denote the assignment where a is paired with d. In our original assignment $\mu_1$, each person is paired with their least preferred roommate, so $\mu_1$ is Pareto dominated by both of the other assignments. Suppose for contradiction that their exists a strategy-proof mechanism M for selecting an efficient, Pareto improving assignment. Note that if a submits the preferences $c \succ b \succ d$ and all other students submit true preferences, then $\mu_2$ is the only assignment that Pareto improves $\mu_1$ (relative to the submitted preferences). In such a case, M must select $\mu_2$. Similarly, if b submits the preferences $c \succ a \succ d$ and all other students submit true preferences, then M must select $\mu_3$ as it is now the only Pareto improving assignment. When all students submit true preferences, M must select either $\mu_2$ or $\mu_3$. If M selects $\mu_2$, then b can do better by deviating and submitting the preferences $c \succ a \succ d$. If M selects $\mu_3$, then a can do better by submitting preferences $c \succ b \succ d$. Either way, M is not strategy proof which is a contradiction. □

## 7.1 Computational Complexity

The purpose of this section is to demonstrate that the Roommate Swap is a polynomial time algorithm and therefore implementable. I demonstrate that it is at worst an $O(N^3)$ algorithm where $N$ is the number of students.

Each iteration of the algorithm involves the following steps, performed in sequence:

1. *Induce the graph.* This is at worst $O(N^2)$ as a graph is defined by its edges and there are at most $\frac{N(N-1)}{2}$ many edges.

2. *Iteratively prune the graph until all blocks are closed under roommates.* West (2001), pg. 157, details an $O(N)$ algorithm for determining blocks. We need to iterate at most $N$ times as each iteration eliminates at least one student from each block or stops looking at a block if it is already closed under roommates. Therefore iteratively pruning the graph is at worst $O(N^2)$.

3. *Find an alternating-cycle.* This process is $O(N)$. At each step we either travel to a previously unvisited vertex, which we can do at most N times, or contract a minimum of two vertices, which we can do at most $\frac{N}{2}$ times. So the algorithm must conclude in at most $N + \frac{N}{2}$ steps. As it takes at most N steps to expand a cycle containing super-vertices to a proper cycle, finding an alternating-cycle concludes in $O(N)$ time.

Therefore each iteration is $O(N^2)$.

**Observation 2.** *In each iteration of the Roommate Swap, at least one student is reassigned her top achievable student.*

*Proof.* The search process ends with an alternating-cycle that may or may not contain super-vertices. Dashed edges from standard vertices are chosen to be the vertex's most preferred student among those who prefer her to their current assignment. Therefore, if the alternating cycle contains no super-vertices, then half the students receive their top achievable match. A grey edge from a super-vertex is not necessarily the student's most preferred achievable student. However, if the alternating-cycle contains a super-vertex and we need to expand our contractions, then there must be a *last* odd-cycle that needs to be expanded.

Figure 6: An odd cycle.

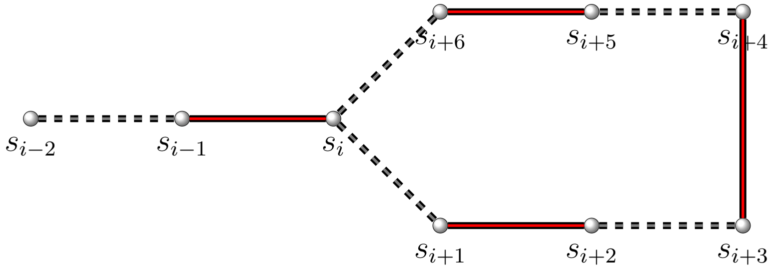


Figure 6 shows a last cycle with six vertices, but the analysis is the same for fewer or greater vertices. Our alternating path must go through $s_i$ and either $s_{i+1}$ or $s_{i+6}$. None of these edges involve super-vertices (this is our last expansion) so by construction, $s_{i+1}$ is $s_i$'s top achievable student and $s_i$ is $s_{i+6}$'s top achievable choice. Either way, at least one student receives her top achievable choice.

□

The significance of this is that once a student has been assigned her top achievable choice, neither she nor her roommate can ever be involved in another Pareto improvement. Therefore we can eliminate them both from consideration. Since we eliminate at least two students after every iteration, there can be at most $\frac{N}{2}$ iterations.

The algorithm performs $O(N)$ many iterations of an $O(N^2)$ process. Therefore it is, at worst, $O(N^3)$.

# References

[1] Abdulkadiroglu, A. and Sonmez, T. (1998), "Random Serial Dictatorship and the Core from Random Endowments in House Allocation Problems," *Econometrica* 66: 689-701.

[2] Chung, K. (2000), "On the Existence of Stable Roommate Matchings," *Games and Economic Behavior* 33: 206-230.

[3] Edmunds, J. (1965), "Paths, Trees, and Flowers," *Canad. J. Math.* 17: 449-467.

[4] Gale, D. and Shapley, L. (1962), "College Admissions and the Stability of Marriage," *Amer. Math. Monthly* 69: 9-15.

[5] Gusfield, D. and Irving, R. (1989), "The Stable Marriage Problem: Structure and Algorithms," MIT Press, Boston, MA.

[6] Roth, A. E., and Sotomayor, M. (1990). "Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis," Econometric Society Monograph 18, Cambridge Univ. Press, Cambridge.

[7] Shapley, L.and Scarf, H. ( 1974), "On Cores and Indivisibility," *Journal of Mathematical Economics* 1: 23-28.

[8] Tan, J. J. M. (1991). A Necessary and Sufficient Condition for the Existence of a Complete Stable Matching, *J. Algorithms* 12: 154178.

[9] West, D. B. (2001). "Introduction to Graph Theory," Prentice Hall, Englewood Cliffs.