FutureMatch: Combining Human Value Judgments and Machine Learning to Match in Dynamic Environments

John P. Dickerson
Carnegie Mellon University
dickerson@cs.cmu.edu

Tuomas Sandholm
Carnegie Mellon University
sandholm@cs.cmu.edu

Abstract

The preferred treatment for kidney failure is a transplant; however, demand for donor kidneys far outstrips supply. *Kidney exchange*, an innovation where willing but incompatible patient-donor pairs can exchange organs—via barter cycles and altruist-initiated chains—provides a life-saving alternative. Typically, fielded exchanges act *myopically*, considering only the current pool of pairs when planning the cycles and chains. Yet kidney exchange is inherently dynamic, with participants arriving and departing. Also, many planned exchange transplants do not go to surgery due to various failures. So, it is important to consider the future when matching.

Motivated by our experience running the computational side of a large nationwide kidney exchange, we present FUTUREMATCH, a framework for *learning to match in a general dynamic model*. FUTUREMATCH takes as input a high-level objective (e.g., "maximize graft survival of transplants over time") decided on by experts, then automatically (i) learns based on data how to make this objective concrete and (ii) learns the "means" to accomplish this goal—a task, in our experience, that humans handle poorly. It uses data from all live kidney transplants in the US since 1987 to learn the quality of each possible match; it then learns the *potentials* of elements of the current input graph offline (e.g., potentials of pairs based on features such as donor and patient blood types), translates these to weights, and performs a computationally feasible batch matching that incorporates dynamic, failure-aware considerations through the weights.

We validate FUTUREMATCH on real fielded exchange data. It results in higher values of the objective. Furthermore, even under economically inefficient objectives that enforce equity, it yields better solutions for the *efficient* objective (which does not incorporate equity) than traditional myopic matching that uses the efficiency objective.

Introduction

Chronic kidney disease is a life-threatening health issue that affects millions of people worldwide; its societal burden is likened to that of diabetes (Neuen et al. 2013). Damage from kidney disease can cause irreparable loss of organ function and, eventually, kidney failure. Such failure requires either continual dialysis or an organ transplant to sustain life.

The preferred treatment for kidney failure is transplantation. However, the demand for donor kidneys is far greater than supply. In the US alone, the waiting list for a kidney transplant had 101,170 patients as of September 8, 2014 (UNOS). Demand is increasing: for example, 36,395 people were added to the US national waiting list in 2013, while only 16,461 left it due to receiving a kidney.

Patients can receive a transplant organ from either a deceased or living donor. Roughly two thirds of transplanted kidneys are sourced from cadavers, while one third come from willing and healthy living donors. Patients who are fortunate enough to find a willing living donor must still contend with *compatibility* issues like blood type, tissue type, and other medical or logistical factors (as we discuss later in the paper). If a willing would-be donor is incompatible with a patient, the kidney cannot be transplanted.

Kidney exchange is a recent innovation that allows patients to swap willing but incompatible donors. Figure 1 shows a graphical representation of a small pool consisting of three patient-donor pairs, where an arrow from pair i to pair j means the patient at j is compatible with the donor at i. Also shown is an altruistic donor; such donors do not come with paired patients and are willing to donate a kidney without asking for one in return. The basic kidney exchange problem is then to recommend an optimal—according to some social welfare function—set of of disjoint cycles and altruist-initiated chains in the graph.

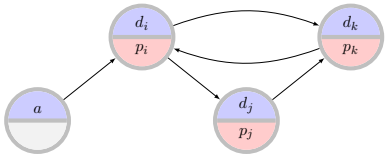

Figure 1: Tiny example kidney exchange pool with three patient-donor pairs and one altruistic donor.

Our Contributions

Fielded kidney exchange algorithms match patients and donors myopically, considering only the current state of the pool when deciding on a matching. However, kidney exchange is inherently dynamic, with patient-donor pairs and altruists arriving and departing the pool over time. Intuitively, the matching algorithm should take distributional information about the future into account when deciding what action to take now. Indeed, recent research has shown that such *dynamic* matching should result in increased overall social welfare (Awasthi and Sandholm 2009; Ünver 2010; Dickerson, Procaccia, and Sandholm 2012; Ashlagi, Jaillet, and Manshadi 2013; Anshelevich et al. 2013; Akbarpour, Li, and Gharan 2014; Anderson et al. 2014). However, most of that prior work has been either

not computationally scalable or studied in simplified theoretical models, while here we study the problem in its full complexity and develop a computationally scalable methodology. Also, the prior work has focused mainly on maximizing the number of matches, while our approach is for any objective. Furthermore, many planned exchange transplants do not go to surgery due to various failures in the last few weeks before the operation (Ashlagi et al. 2011; Dickerson, Procaccia, and Sandholm 2013). That introduces a second form of dynamism. *The approach in this paper is the first to address both forms of dynamism.*

As a proof of concept for this paper, we learn a novel transplant quality predictor from a dataset consisting of all living-donor kidney transplant events in the US since Oct 1, 1987, and discover that some present-day features of living-donor transplants do not align with older results from deceased donation in the medical literature (Opelz 1985). This is noteworthy since today's exchange priority policies have largely been inherited from the United Network for Organ Sharing (UNOS) deceased-donor waiting list policies.

Motivated by our experience running the computational side of a large nationwide kidney exchange that includes over 130 transplant centers (59% of the centers in the US), we present FUTUREMATCH, a general framework for learning how to match in dynamic environments. FUTUREMATCH separates the "means" from the "ends" of kidney exchange; it takes as input from human experts an overarching objective, and automatically learns a matching strategy to achieve this goal. We validate the framework on three example objective functions on real data drawn from the large, fielded exchange. We find that using FUTUREMATCH *even with economically inefficient objectives*—like maximizing the match size subject to equity constraints—results in significantly higher efficiency than myopic matching with the explicit objective of efficiency.

Kidney Exchange Model

The standard model for kidney exchange encodes a kidney exchange as a directed *compatibility graph* $G = (V, E)$ by constructing one vertex for each patient-donor pair in the pool (Roth, Sönmez, and Ünver 2004; 2005a; Roth, Sönmez, and Ünver 2005b). An edge e from v_i to v_j is included if the patient in v_j wants the donor kidney of v_i. A paired donor is willing to give her kidney if and only if the patient in her vertex v_i receives a kidney.

The weight w_e of an edge e represents the utility to v_j of obtaining v_i's donor kidney. A cycle c in the graph G represents a possible kidney swap, with each vertex in the cycle obtaining the kidney of the previous vertex. If c includes k pairs, we refer to it as a *k-cycle*. For example, the compatibility graph in Figure 1 includes two possible cycles: a 2-cycle between vertex v_i and v_k, and a 3-cycle consisting of vertices v_i, v_j, and v_k. In kidney exchange, cycles of length at most some small constant L are allowed—all transplants in a cycle must be performed simultaneously so that no donor backs out after his patient has received a kidney but before he has donated his kidney. In most fielded kidney exchanges, $L = 3$ (i.e., only 2- and 3-cycles are allowed).

Fielded kidney exchanges gain great utility through the use of chains (Montgomery et al. 2006; Rees et al. 2009). Chains start with an altruistic donor donating his kidney to a patient, whose paired donor donates her kidney to another patient, and so on. The compatibility graph in Figure 1 includes four possible chains: $\langle a, v_i \rangle$, $\langle a, v_i, v_j \rangle$, $\langle a, v_i, v_k \rangle$, and $\langle a, v_i, v_j, v_k \rangle$. Chains can be (and typically are) longer than cycles in practice because it is not necessary to carry out all the transplants in a chain simultaneously. There is a chance that a donor backs out of his/her commitment to donate—which has happened (albeit rarely) already in the United States. Cycles cannot be executed in parts because if someone backs out of a cycle, then some pair has lost a kidney (i.e., their "bargaining chip"). In contrast, if someone backs out of a chain, no pair has lost their bargaining chip (although it is unfortunate that the chain ends).

A *matching* M is a collection of disjoint cycles and chains in the graph G. The cycles and chains must be disjoint because no donor can give more than one of her kidneys. Given the set of all legal matchings $\mathcal{M}$, the (batch) *clearing problem* in kidney exchange is to find a matching M^* that maximizes some utility function $u : \mathcal{M} \to \mathbb{R}$. Formally:

$$M^* = \arg\max_{M \in \mathcal{M}} u(M)$$

The standard clearing problem for finite cycle cap $L > 2$ (even without chains) is NP-hard (Abraham, Blum, and Sandholm 2007). Abraham, Blum, and Sandholm (2007) took the first serious computational step toward solving the kidney exchange problem by providing a specialized branch-and-price-based (Barnhart et al. 1998) integer program solver; subsequent work by Dickerson, Procaccia, and Sandholm (2013; 2014) has increased solver speed and generality. We use an adapted version of that clearing algorithm as the batch clearing algorithm module in our framework (as we will discuss later).

In fielded kidney exchanges, one typically finds the maximum weighted cycle cover (i.e., $u(M) = \sum_{c \in M} \sum_{e \in c} w_e$). This *utilitarian* objective can favor certain classes of patient-donor pairs while marginalizing others, a phenomenon that we explore—and help alleviate—in this paper.

Proposed Method

We now present the FUTUREMATCH framework for learning to match in dynamic environments. We begin by motivating and describing the framework at a high level in the first subsection. In the following subsections we discuss the different parts of the framework in detail and how we instantiated them for kidney exchange.

The FUTUREMATCH Framework

We are interested in learning from demographically accurate data how to match *in the present* such that some overarching objective function is maximized over time. Scalability is important: heavy offline statistics can be computed and periodically updated, but the fielded clearing algorithm must run quickly (within minutes or at most hours).

Figure 2 graphically depicts the FUTUREMATCH framework. A domain expert (e.g., a committee of medical and legal professionals) begins by describing an overall objective function for the exchange. Even *measuring* this objective can be difficult: for example, if the goal is to maximize the number of days added to patients' lives via kidney transplantation, then calculating the relative quality of a proposed match requires knowing some notion of utility for each edge—representing a potential transplant—in the

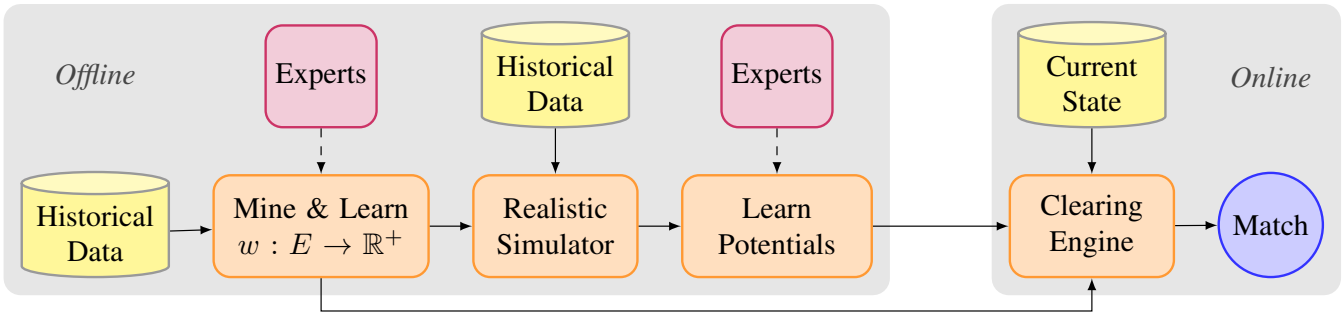

Figure 2: The FUTUREMATCH framework.

compatibility graph. We propose to learn this edge weight function $w : E \to \mathbb{R}^+$ from data, and give examples for a variety of objective functions later.

The learned weight function w is then fed into a parameterized (kidney exchange) simulator, calibrated by real data so that it mimics the underlying distribution. This generator in turn feeds training and test sets into a system for learning the *potentials* of element classes in the compatibility graph. Intuitively, given an element θ (e.g., vertex, edge, cycle, or chain type), a potential $P_\theta \in \mathbb{R}$ quantifies the expected utility to the exchange of that element in the future (Dickerson, Procaccia, and Sandholm 2012). Potentials are combined with w to quantify an edge-specific quality rating. Here, we learn potentials for the combinations of different blood types for pairs under each of the weight functions we define.

Finally, the fielded clearing algorithm incorporates the combined weight function w and set of potentials $\mathcal{P}_\Theta$ into its myopic weighted matching algorithm. This incorporation comes at very low or no cost to the runtime of the clearing algorithm; indeed, the final "potential-aware" input graph is simply a re-weighted version of the original compatibility graph, using the weights that encode the future.

In the rest of the section, we describe an in-depth implementation of FUTUREMATCH. Our goals are twofold: first, to show the general applicability and tractability of the framework, and second, to mimic a large fielded kidney exchange. Accomplishing this second goal, and leveraging our involvement with fielded kidney exchanges, sets the stage for adoption of sustainability-motivated technology that solves a problem clearly too difficult for humans—a success story for computational sustainability practitioners.

Encoding an Objective Function

We now discuss in depth the process of defining an objective for FUTUREMATCH. We do this in the context of kidney exchange, but note that the process is general.

The medical and legal communities in kidney exchange are concerned about a wide variety of match characteristics. In our experience, the most frequently discussed include the number of overall matches, the number of overall transplants, the quality of transplants, and whether or not to prefer specific subgroups in the exchange (children, sensitized patients, underrepresented ethnicities) and by how much. Other concerns could include notions of fair treatment among participating centers and minimizing legal exposure.

In this paper, we consider two different kidney exchange models—deterministic, where post-algorithmic match failures are not quantified in the optimization problem and failure-aware, where they are—and three matching objectives in each of the two models:

1. MAXCARD: Maximize the total number (i.e., cardinality) of patients who are algorithmically matched (in the deterministic model) or receive transplants in expectation (in the failure-aware model);
2. MAXCARD-FAIR: Maximize the total number of patients who are algorithmically matched (in the deterministic model) or receive transplants in expectation (in the failure-aware model), where "marginalized" patients are weighted in the objective by some constant factor β more than others; and
3. MAXLIFE: Maximize the total time algorithmically-matched (deterministic) or transplanted (failure-aware) donor organs will last in patients.

Each of these objectives amounts to setting weights on edges in the input graph (e.g., Figure 1). Next, we detail edge weighting algorithms for these example objectives.

In our experience, when committees debate priority points for today's exchanges, the discussions confound the goal and the means. For example, a goal could be to maximize matches and a means could be to prioritize sensitized patients because they are harder to match in the future. On the other hand, many argue that sensitized patients should be inherently preferred, and it seems that most do not make a clear distinction between means and ends. In contrast, FUTUREMATCH clearly *separates ends and means*. Our objective (i.e., the "end") lives in the space of weights on edges, which the committee can clearly debate. On the other hand, our framework automatically optimizes, via learning, the potentials (the "means") that are used as the means for enabling the algorithm to make good future-aware failure-aware decisions. The committee does not need to debate these potentials, whose quantitative impact on performance is hard for a human to predict or even understand.

Defining MAXCARD and MAXCARD-FAIR. The MAXCARD-FAIR objective can be viewed as a generalized form of MAXCARD (that is, MAXCARD is just MAXCARD-FAIR with an empty set of vertices who are preferred by the objective). A natural weighted fairness rule, adapted from (Dickerson, Procaccia, and Sandholm 2014), adjusts edge weights by some re-weighting function $\Delta : E \to \mathbb{R}^+$. A simple example re-weighting function is multiplicative:

$$\Delta^\beta(e) = \begin{cases} (1+\beta)w_e & \text{if } e \text{ ends in } V_P \\ w_e & \text{otherwise} \end{cases}$$

Here, $V_P \subseteq V$ is the set of preferred vertices (we will define one such subset in the experiments). Intuitively, for some $\beta > 0$, this function scales the weight of edges ending in marginalized vertices by $(1 + \beta)$. For example, if $\beta = 1.5$, then the optimization algorithm will value edges that result in a marginalized patient receiving a transplant at 250% of their initial weight (possibly scaled by factors such as edge failure probability or chain position, as we discuss later).

For any $M \in \mathcal{M}$, let M' be the matching such that every edge $e \in E$ has augmented weight $\Delta^\beta(e)$. Then the MAX-CARD-FAIR utility function u_Δ is defined in terms of the utilitarian MAXCARD utility function u applied to the augmented matching M', such that $u_\Delta(M) = u(M')$. In the experiments, we vary the parameter β to empirically quantify its effects on each of the three objective functions.

Optimizing for MAXLIFE via learning to predict graft survival from data. With the MAXLIFE objective we are interested in maximizing how long the transplanted kidneys, in aggregate, survive in the patients.[1] To do so, we must first determine an empirically sound estimate of the lifespan of a transplant as a function of donor and recipient attributes.

Delen, Walker, and Kadam (2005) compare a variety of techniques for predicting breast cancer survivability; unlike their study, we are interested in predicting the survival *length* of a kidney graft, as opposed to whether or not a patient survives treatment at all. Data mining models are also actively being developed to predict the risk of readmission for congestive heart failure patients (Meadem et al. 2013). Most related to our work is the Kidney Donor Profile Index (KDPI), which is currently under development by UNOS for use in the deceased donor allocation process (Kidney Transplantation Committee 2011). The KDPI score of a deceased donor kidney measures the estimated quality of the donor organ being allocated to the *average* recipient. In contrast, our predictor, which we will describe next, provides a unique quality score not just based on donor attributes but also based on attributes of the specific potential recipient.

We look at all $75{,}264$ *living donor* transplant events in the US between October 1, 1987 and June 30, 2013. This data includes medical characteristics of the recipient and donor at the time of transplantation, as well as follow-up data regarding the health of the recipient and the recipient's new kidney; this follow-up data is updated at least annually.

Conditioned on a kidney graft being marked as failed in our dataset, the average graft lifetime is about 1912.7 days, or slightly over 5 years. However, due in large part to the marked increase in kidney failure since the late 1980s, nearly 75% of grafts in the dataset are not marked as failed. This occurs because either (i) the recipient is still alive with a functioning donated kidney or (ii) the recipient has died, but for a non-kidney-failure-related reason. Thus, we use *survival analysis* to estimate the lasting power of a graft.

Features of both the recipient and donor have a large effect on graft survival. For example, tissue type (HLA) testing measures the closeness of match between antigens in the cells of a donor and patient. Figure 3 gives a Kaplan-Meier estimator of the survival functions of (i) kidney transplants resulting from a donor and recipient being a perfect HLA match and (ii) those resulting from imperfect HLA matchings. Clearly, a kidney that is a perfect tissue type match is more desirable than an imperfectly matched one; indeed, the model estimates a median survival time of 5808 days for a perfect match compared to 4300 for an imperfect match. A log-rank test revealed that the difference between the two distributions was significant ($p \ll 0.0001$).

In our experiments, we use a Cox proportional hazards regression analysis to explore the effect of multiple features on survivability. At a high level, this method regresses the survival time of the graft against explanatory features of the donor and recipient. More specifically, define the *hazard* H at time t days after a transplant as follows:

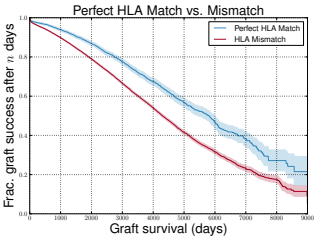

Figure 3: Kaplan-Meier estimator of survival functions, 95% confidence intervals.

$$H(t) = H_0(t) \times \exp(b_1 X_1 + b_2 X_2 + \ldots + b_k X_k) \quad (1)$$

Here, each X_i is a predictor variable corresponding to a single feature of the donor or recipient, and $H_0(t)$ is a baseline hazard rate at time t for a recipient with $X_i = 0$ for $i \in \{1, \ldots, k\}$. Then $H(t)$ represents the instantaneous risk of graft failure at time t. We want to learn this function.

To begin, we include the following features: recipient age, difference in donor and recipient's age, donor HLA profile, recipient HLA profile, donor and recipient blood type compatibility. The HLA profile of a donor or recipient is separated into three integral features—HLA-A, HLA-B, and HLA-DR—that can take values in $\{0, 1, 2\}$, representing 0, 1, or 2 mismatches. By separating the general HLA mismatch feature into three separate mismatch features, we complicate (but increase the power of) the model (Opelz 1985); this separation is motivated by evidence that mismatches at the HLA-A, -B, and -DR level have varyingly negative impact on survival.

We ran a Cox proportional hazards regression on this unpruned feature space. This used $74{,}244$ live donor transplantations during which there were $18{,}714$ graft failures (920 live donation events were dropped due to one or more missing features). Our initial regression showed that increases in the HLA-B mismatch feature did not have a significant effect on the dependent variable ($p = 0.22$). Prior research from the mid-1980s on *cadaveric* donation found a significant relationship between the combined feature of HLA-B and HLA-DR mismatches on graft outcome (Opelz 1985); we find that this does not hold on living donor data in the present. After selecting significant variables in this initial run—that is, all of the attributes previously discussed *except* HLA-B mismatch—we re-ran a Cox proportional hazard regression. Results are reported in Table 1.

Table 1 gives the standard error, z-score, and corresponding p-value for each of our pruned features; each clearly has a statistically significant effect on graft survival. To interpret the results, as an example first consider the HLA-DR feature. We see that $\exp(b_{\text{HLA-DR}}) \approx 1.087$; recalling Equation 1, a unit increase in the HLA-DR mismatch feature will

[1] Another objective would be to maximize aggregate increase in life duration. This would involve subtracting out the expected life duration without a transplant from the expected life duration with the transplant, and could incorporate the possibility of additional transplants after graft failure.

feature	$\exp(b_i)$	$\text{SE}(b_i)$	z	p
recipient age	1.00753	0.0008	9.715	$< 2 \times 10^{-16}$
age diff.	1.00525	0.0007	7.766	8.10×10^{-15}
HLA-A	1.05273	0.0120	4.297	1.73×10^{-5}
HLA-DR	1.08680	0.0119	6.984	2.86×10^{-12}
ABO incomp.	1.37871	0.0748	4.295	1.74×10^{-5}

Table 1: Learned weights via Cox regression after feature pruning for statistical significance.

result in a factor of 1.087 increase over the baseline hazard rate. Varying either recipient age or the difference in donor and recipient age was also statistically significant, with a unit increase in recipient age having a larger effect on the hazard rate. As might be expected, blood type compatibility plays the largest role in hazard rate, where an incompatible (and thus heavily immunosuppressed) transplant has a factor 1.379 increase over the base hazard rate.

Using this data, we can estimate $S_e(t)$, the survival probability at time t for a potential transplant $e \in E$ between a recipient and donor with features x_i^e, as follows: $S_e(t) = \exp\left(-H_0(t) \times \sum_i x_i^e b_i\right)$. Building on this, we define a weight function $w : E \to \mathbb{R}^+$ as $w(e) \propto \exp\left(-\sum_i x_i^e b_i\right)$. Intuitively, the weight function w assigns higher relative weight to edges with lower risk, in turn biasing the optimizer toward transplants with longer expected graft survival.

Learning the Potentials

The weight function w defined above quantifies how useful an edge is *in the present*. To complement this, *potentials* quantify how useful an element of the graph (an edge, 2-cycle, etc.) is expected to be in the future. This idea was initially proposed by Dickerson, Procaccia, and Sandholm (2012); we build on their work but use a better learning algorithm (theirs did not converge, while ours did) for determining the potentials, and a significantly more realistic distribution of training and test instances.

First, we select a set Θ of features representing different element types in the pool. Then, for each element type θ, assign some value $P_\theta \in \mathbb{R}$ that represents the expected potential usefulness of that kind of element to the pool over time. (We give a pedagogical instantiation of Θ in Appendix D.) In this paper, we use a rich feature space consisting of the *ABO blood types* of patients and donors. The blood type of a donor kidney can result in rejection in a potential patient. Human blood is split into four types—O, A, B, and AB—based on the presence or absence of the A and B proteins. While other reasons for incompatibility also exist (e.g., due to sensitization, sickness, etc.), a type O kidney can be transplanted into any patient; type A and B kidneys can be transplanted into A and B patients, respectively, or an AB candidate; and type AB kidneys are limited to only type AB patients. This imposes a natural partial ordering on blood types; for example, it seems that an O-donor is somehow more "valuable" than an A-donor because she has no blood type restriction. Our automatically learned potentials agree with this, as we will now discuss.

There are $4 \times 4 = 16$ combinations of patient and donor blood types, and 4 possible blood types of altruists. So, we have 20 different kinds of vertices. We want to learn a potential for each of those 20 vertex types. Formally, the types of vertex are $\Theta_{\text{ABO}} = \{\text{O-O}, \text{O-A}, \ldots, \text{AB-B}, \text{AB-AB}\} \cup \{\text{O}, \ldots, \text{AB}\}$, and we want to learn values P_θ for $\theta \in \Theta_{\text{ABO}}$.

We combine the learned potentials P_θ with the weight function w learned earlier using a function $f_w : E \to \mathbb{R}$. It balances the *myopic* value of an edge encoded by w with the *future value* of an edge encoded by potentials. The idea is that the revised weight, f_w, of an edge is its immediate value if matched minus the potentials of its vertices because if those vertices are matched now, they cannot contribute to future matches. Specifically, $f_w(e) = w(e) \cdot (1 - P_{\theta_d} - P_{\theta_p})$, where d and p are the vertices adjacent to edge e.

We use SMAC (Hutter, Hoos, and Leyton-Brown 2011), a state-of-the-art model-based algorithm configuration tool that searches through a parameter space to optimize a given objective. SMAC guides its navigation in the space of parameter vectors by constructing a model that predicts algorithm performance as a function of the parameters. It uses the model to select promising parameter vectors and tests them against the incumbent parameter vector.

In our setting, the parameter vector is the vector of potentials. At each parameter vector that SMAC navigates to, we run a large number of trials of our simulator of a kidney exchange to see how the batch matching algorithm would perform in a dynamic setting using that parameter vector. That performance number is then fed back to SMAC, and SMAC navigates to the next parameter vector to continue the search. We learned potentials in two models—*deterministic*, where post-algorithmic match failures are not quantified in the optimization problem and *failure-aware*, where they are—using a realistic dynamic simulator we built based on historical data from the UNOS kidney exchange (Kidney Paired Donation Work Group 2013). See Appendix E for details.

Experiments

We validated FUTUREMATCH experimentally on data from the UNOS nationwide kidney exchange. We explore the effect each of the three objectives—MAXCARD, MAXCARD-FAIR, and MAXLIFE—has on a variety of metrics under FUTUREMATCH and under myopic deterministic matching, which is the fielded state of the art. The latter does not take edge failure or learned potentials into account during optimization; as described earlier, it finds a maximum weight matching (i.e., for each chain or cycle c, $u(c) = \sum_{e \in c} w_e$) during each period separately.

In the fairness-weighted expariments, we adapt our matching algorithm using the re-weighting function Δ^β described earlier. The preferred set of vertices V_P includes those with a pediatric or highly-sensitized patient. These preferences are commonly used in kidney exchanges, albeit not in sophisticated, quantitative ways. For kidney exchange it has explicitly been articulated that pediatric patients should be preferred not only because they have a lot of life left (barring their kidney disease) but also because having poor kidney function stunts growth. Similarly, some patients are *highly sensitized*, which means they are extremely unlikely to be medically compatible with a random organ. For these patients, finding a kidney is difficult (UNOS). The percentage of highly-sensitized patients in fielded kidney exchanges is high; over 60% of the patients in the UNOS kidney exchange are highly sensitized. In fielded exchanges, both of these "marginalized" patient types are prioritized.

Total	$\|V\| = 300$		$\|V\| = 400$		$\|V\| = 500$		$\|V\| = 600$		$\|V\| = 700$		$\|V\| = 800$		$\|V\| = 900$	
	Gain	p	*Gain*	p	*Gain*	p	*Gain*	p	*Gain*	p	*Gain*	p	*Gain*	p
MAXCARD	+2	✓	+4	✓	+5	✓	+6	✓	+10	✓	+11	✓	+13	✓
MAXCARD-FAIR, $\beta = 1$	+1	✓	+4	✓	+6	✓	+8	✓	+9	✓	+11	✓	+12	✓
MAXCARD-FAIR, $\beta = 2$	+1		+2	✓	+3	✓	+3	✓	+5	✓	+6	✓	+10	✓
MAXCARD-FAIR, $\beta = 3$	+1		+0		+3	✓	+1		+1	✓	+3	✓	+2	
MAXCARD-FAIR, $\beta = 4$	-1		+1		+1		+1		+3	✓	+3		+2	
MAXCARD-FAIR, $\beta = 5$	+0		+0		+1		+1		+1		+2		+3	
MAXLIFE	+2	✓	+3	✓	+6	✓	+8	✓	+7	✓	+11	✓	+9	✓
Marginalized														
MAXCARD	-2	✗	-2	✗	-3	✗	-4	✗	-6	✗	-7	✗	-9	✗
MAXCARD-FAIR, $\beta = 1$	-1	✗	-1	✗	-1	✗	-2	✗	-3	✗	-3	✗	-5	✗
MAXCARD-FAIR, $\beta = 2$	+0		+0		+1	✓	+1	✓	+2	✓	+1		+1	
MAXCARD-FAIR, $\beta = 3$	+1	✓	+1	✓	+3	✓	+3	✓	+3	✓	+5	✓	+4	✓
MAXCARD-FAIR, $\beta = 4$	+1	✓	+2	✓	+3	✓	+4	✓	+4	✓	+5	✓	+5	✓
MAXCARD-FAIR, $\beta = 5$	+1	✓	+2	✓	+3	✓	+4	✓	+5	✓	+7	✓	+5	✓
MAXLIFE	-1	✗	-3	✗	-3	✗	-5	✗	-6	✗	-6	✗	-9	✗

Table 2: Median gains in expected total number of transplants (top table) and total number of marginalized transplants (bottom table) under FUTUREMATCH. A ✓ represents statistical significance (Wilcoxon signed-rank test, $p \ll 0.01$).

We quantitatively explore how this should be done and what the impact is.

Results. We compare FUTUREMATCH against a baseline of myopic deterministic matching under each of the objectives. Conservatively, statistical significance was determined using the Wilcoxon signed-rank test, which is a nonparametric alternative to the paired t-test. Table 2 shows the median expected gain in the overall number of transplants from using FUTUREMATCH under each of the objectives. Each column labeled $|V| = k$ corresponds to a simulation over k patient-donor pairs and altruists; we test over increasing values of k because kidney exchanges (both in the US and worldwide) are still expanding toward their steady-state sizes.

Table 2 (top) shows that the two objectives that do not regard fairness—MAXCARD and MAXLIFE—significantly beat myopic deterministic matching under the same objective. Interestingly, so too does MAXCARD-FAIR for low values of β. As β increases, the gain in *overall* number of transplants decreases (although it never drops below the deterministic matching algorithm with significance). This decrease in overall gain is incurred because marginalized patients, who (i) generally have lower in-degree, and (ii) have a higher probability of match failure, are being weighted more than easier-to-match pairs.

Table 2 (bottom) explores this tradeoff between fairness and efficiency explicitly. For the fairness-agnostic and lightly fairness-preferring objectives, a relative loss of a few marginalized transplants is realized—although this loss of marginalized transplants is always less (typically much less) than the overall gain in transplants. Increasing the optimizer's preference for marginalized patients results in statistically significant gains in the number of marginalized transplants at *no* statistically significant loss in the overall expected number of transplants. In fact, for a middle ground around $\beta = 2$, FUTUREMATCH often shows statistically significant gains in *both* overall transplant and marginalized transplant counts—a clear win over myopia.

Our experiments support the following conclusions:

- FUTUREMATCH under MAXCARD and MAXCARD-FAIR with low $\beta = 1$ results in a significant increase in the *overall* number of transplants compared to myopic, at the cost of a smaller decrease in the number of *marginalized* transplants.
- FUTUREMATCH under MAXCARD-FAIR with high β results in a significant increase in *marginalized* transplants, at *no cost* to the overall number of transplants under myopic matching.
- For a middle ground around $\beta = 2$, FUTUREMATCH can result in both more overall expected transplants and more marginalized transplants.

We note that we are *not* making policy recommendations; rather, we are giving a proof of concept that our framework can effectively balance conflicting wants in an exchange. Indeed, the exact fairness quantification β that most effectively balances efficiency and fairness is a function of the underlying graph dynamics, which vertices are considered marginalized, and the ethical and legal wants of an exchange. All of these dimensions can be effectively encoded, validated, compared, and fielded through FUTUREMATCH.

Conclusions

We presented FUTUREMATCH, a framework for learning to do complex matching in a general dynamic model. The framework addresses a computational sustainability problem directly uses data mining and optimization in many of its modules. Motivated by our experience running the computational side of a large nationwide kidney exchange, we showed how to instantiate FUTUREMATCH to mimic an exchange under three different matching objectives and under two models of kidney exchange. We validated FUTUREMATCH on real data drawn from 94 match runs of the US nationwide exchange, and found that dynamic matching results in statistically significant increases in each of these objectives. Perhaps most critically, we showed that the framework yields better efficiency *and* better fairness than deterministic myopic matching algorithms—which are the status quo class of algorithm in practice.

References

Abraham, D.; Blum, A.; and Sandholm, T. 2007. Clearing algorithms for barter exchange markets: Enabling nationwide kidney exchanges. In *Proceedings of the ACM Conference on Electronic Commerce (EC)*, 295–304.

Akbarpour, M.; Li, S.; and Gharan, S. O. 2014. Dynamic matching market design. In *Proceedings of the ACM Conference on Economics and Computation (EC)*.

Anderson, R.; Ashlagi, I.; Gamarnik, D.; and Kanoria, Y. 2014. A dynamic model of barter exchange. Working paper.

Anshelevich, E.; Chhabra, M.; Das, S.; and Gerrior, M. 2013. On the social welfare of mechanisms for repeated batch matching. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 60–66.

Ashlagi, I.; Gilchrist, D. S.; Roth, A. E.; and Rees, M. 2011. Nonsimultaneous chains and dominos in kidney-paired donation—revisited. *American Journal of Transplantation* 11(5):984–994.

Ashlagi, I.; Jaillet, P.; and Manshadi, V. H. 2013. Kidney exchange in dynamic sparse heterogenous pools. In *Proceedings of the ACM Conference on Electronic Commerce (EC)*, 25–26.

Awasthi, P., and Sandholm, T. 2009. Online stochastic optimization in the large: Application to kidney exchange. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, 405–411.

Barnhart, C.; Johnson, E. L.; Nemhauser, G. L.; Savelsbergh, M. W. P.; and Vance, P. H. 1998. Branch-and-price: Column generation for solving huge integer programs. *Operations Research* 46(3):316–329.

Blum, A.; Gupta, A.; Procaccia, A. D.; and Sharma, A. 2013. Harnessing the power of two crossmatches. In *Proceedings of the ACM Conference on Electronic Commerce (EC)*, 123–140.

Delen, D.; Walker, G.; and Kadam, A. 2005. Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine* 34(2):113–127.

Dickerson, J. P.; Procaccia, A. D.; and Sandholm, T. 2012. Dynamic matching via weighted myopia with application to kidney exchange. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 1340–1346.

Dickerson, J. P.; Procaccia, A. D.; and Sandholm, T. 2013. Failure-aware kidney exchange. In *Proceedings of the ACM Conference on Electronic Commerce (EC)*, 323–340.

Dickerson, J. P.; Procaccia, A. D.; and Sandholm, T. 2014. Price of fairness in kidney exchange. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

Feldman, J.; Mehta, A.; Mirrokni, V.; and Muthukrishnan, S. 2009. Online stochastic matching: Beating 1-1/e. In *Symposium on the Foundations of Computer Science (FOCS)*.

Goel, G., and Tripathi, P. 2012. Matching with our eyes closed. In *Symposium on the Foundations of Computer Science (FOCS)*, 718–727. IEEE.

Hutter, F.; Hoos, H.; and Leyton-Brown, K. 2011. Sequential model-based optimization for general algorithm configuration. In *Proc. of LION-5*, 507–523.

Karp, R. M.; Vazirani, U. V.; and Vazirani, V. V. 1990. An optimal algorithm for on-line bipartite matching. In *Proceedings of the Annual Symposium on Theory of Computing (STOC)*, 352–358.

Kidney Paired Donation Work Group. 2013. OPTN KPD pilot program cumulative match report (CMR) for KPD match runs: Oct 27, 2010 – Apr 15, 2013.

Kidney Transplantation Committee. 2011. OPTN concepts for kidney allocation.

Manshadi, V. H.; Gharan, S. O.; and Saberi, A. 2012. Online stochastic matching: Online actions based on offline statistics. *Mathematics of Operations Research* 37(4):559–573.

Meadem, N.; Verbiest, N.; Zolfaghar, K.; Agarwal, J.; Chin, S.-C.; and Roy, S. B. 2013. Exploring preprocessing techniques for prediction of risk of readmission for congestive heart failure patients. In *Data Mining and Healthcare (DMH), at International Conference on Knowledge Discovery and Data Mining (KDD)*.

Molinaro, M., and Ravi, R. 2013. Kidney exchanges and the query-commit problem. Manuscript.

Montgomery, R.; Gentry, S.; Marks, W. H.; Warren, D. S.; Hiller, J.; Houp, J.; Zachary, A. A.; Melancon, J. K.; Maley, W. R.; Rabb, H.; Simpkins, C.; and Segev, D. L. 2006. Domino paired kidney donation: a strategy to make best use of live non-directed donation. *The Lancet* 368(9533):419–421.

Neuen, B. L.; Taylor, G. E.; Demaio, A. R.; and Perkovic, V. 2013. Global kidney disease. *The Lancet* 382(9900):1243.

Opelz, G. 1985. Correlation of HLA matching with kidney graft survival in patients with or without cyclosporine treatment: for the collaborative transplant study. *Transplantation* 40(3):240–242.

Rapaport, F. T. 1986. The case for a living emotionally related international kidney donor exchange registry. *Transplantation Proceedings* 18:5–9.

Rees, M.; Kopke, J.; Pelletier, R.; Segev, D.; Rutter, M.; Fabrega, A.; Rogers, J.; Pankewycz, O.; Hiller, J.; Roth, A.; Sandholm, T.; Ünver, U.; and Montgomery, R. 2009. A nonsimultaneous, extended, altruistic-donor chain. *New England Journal of Medicine* 360(11):1096–1101.

Roth, A.; Sönmez, T.; Ünver, U.; Delmonico, F.; and Saidman, S. L. 2006. Utilizing list exchange and nondirected donation through 'chain' paired kidney donations. *American Journal of Transplantation* 6:2694–2705.

Roth, A.; Sönmez, T.; and Ünver, U. 2004. Kidney exchange. *Quarterly Journal of Economics* 119(2):457–488.

Roth, A.; Sönmez, T.; and Ünver, U. 2005a. A kidney exchange clearinghouse in New England. *American Economic Review* 95(2):376–380.

Roth, A.; Sönmez, T.; and Ünver, U. 2005b. Pairwise kidney exchange. *Journal of Economic Theory* 125(2):151–188.

Saidman, S. L.; Roth, A.; Sönmez, T.; Ünver, U.; and Delmonico, F. 2006. Increasing the opportunity of live kidney donation by matching for two and three way exchanges. *Transplantation* 81(5):773–782.

United Network for Organ Sharing (UNOS). http://www.unos.org/.

Ünver, U. 2010. Dynamic kidney exchange. *Review of Economic Studies* 77(1):372–414.

Appendix A: Additional Prior Research in Dynamic Matching

We briefly overview related theoretical work in dynamic matching. Our work differs from it significantly. For one, FUTUREMATCH learns to match better using data. Also, it applies to a significantly richer set of problems, such as kidney exchange, than traditional matching where the goal is to simply pair up vertices. We now briefly overview some of the prior work.

In the dynamic matching problem, vertices and/or edges arrive and depart over time. Karp, Vazirani, and Vazirani showed that the competitive ratio of any randomized online bipartite matching algorithm is at best $1 - 1/e$, and gave an algorithm that achieves this (Karp, Vazirani, and Vazirani 1990). The online bipartite matching problem has seen significant renewed interest from Internet advertising firms, since it relates to keyword and display ad allocation (e.g., (Feldman et al. 2009; Manshadi, Gharan, and Saberi 2012)). Akbarpour, Li, and Gharan (2014) looks at minimizing a function of average vertex waiting time (instead of overall number of matches) in a dynamic model where vertices arrive and depart over time.

In the *query-commit* problem (Goel and Tripathi 2012; Molinaro and Ravi 2013), the goal is to find a matching of maximum size in a graph where the set of edges is not known ahead of time. Instead, an actor may *query* an edge and, if present, is forced to *commit* to using that edge in the final matching. This is relevant to kidney exchange, where a matching algorithm first selects an edge, but that edge can fail for either a medical or logistical reason before transplantation. Blum et al. (2013) address a similar problem, where at most two edges incident to any vertex can be selected in a possible matching. Those papers operate in significantly simpler models than real kidney exchange.

Dickerson, Procaccia, and Sandholm (2013) show that making the optimization "failure-aware" by incorporating edge failure probabilities directly into the optimization process increases the expected number of transplants both in theory and in practice; we incorporate their (static) model into the FUTUREMATCH framework. Dickerson, Procaccia, and Sandholm (2014) also explored the *price of fairness* in kidney exchange, which is a measure of the relative loss in efficiency of the system under various "fair" matching rules. They found that while the theoretical price of fairness is low, in practice it can be non-trivial; we also incorporate this model into FUTUREMATCH (and show that a favorable balance between fairness and efficiency can be struck in practice).

Dynamic kidney exchange is a largely unsolved problem. From the theory side, Ünver provided the first results in a model of dynamic kidney exchange that only includes 2-cycles and no chains (Ünver 2010). Ashlagi, Jaillet, and Manshadi (2013) and Anshelevich et al. (2013) look at the *batch matching* problem, where a maximum cardinality matching is performed every k time periods, or whenever some feature of the graph crosses a predetermined threshold. Awasthi and Sandholm (2009) approached the problem computationally, using trajectory-based optimization that samples potential future states to inform the present matching algorithm; unfortunately, this does not scale computationally to the projected steady-state size of the nationwide kidney exchange. Motivated by this, Dickerson, Procaccia, and Sandholm (2012) proposed learning offline the *potentials* of different elements (e.g., types of vertices or edges) in the input graph, then subtracting out these potentials per element in the objective online. We incorporate this approach, too, into FUTUREMATCH.

Explicit novelty of our optimization approach. This paper's high level contribution is the design, implementation, and validation of FUTUREMATCH, the first data-driven learning framework for complex online matching. *This paper is the first of its kind to combine future-aware (Dickerson, Procaccia, and Sandholm 2012), failure-aware (Dickerson, Procaccia, and Sandholm 2013), and fairness-aware (Dickerson, Procaccia, and Sandholm 2014) dimensions explicitly in the optimization and simulation models.* Each of these points has been discussed in the kidney exchange literature and—especially in the context of fairness—continues to be discussed in fielded kidney exchanges. We provide the first computational framework in which the tradeoffs and benefits of these aspects of kidney exchange can be explored, quantified, and balanced to maximize a high-level objective provided by human experts.

Appendix B: Fielded Kidney Exchanges

The idea of kidney exchange was presented in 1986 (Rapaport 1986), while the first organized kidney exchange, the New England Paired Kidney Exchange (NEPKE), started in 2003–2004 (Roth, Sönmez, and Ünver 2004; 2005a; Roth, Sönmez, and Ünver 2005b). It has since ceased operations and its pool was merged into the UNOS kidney exchange, which started in 2010 and now includes 58% of the US transplant centers. All the decisions are transparent and purely computational without human intervention.

There are also two large private kidney exchanges in the US, the National Kidney Registry (NKR) and the Alliance for Paired Donation (APD). They typically only work with large transplant centers. Transplant centers can be part of multiple exchanges. NKR makes their matching decisions manually and APD uses a combination of algorithmic and manual decision making. There was also another large private kidney exchange, the Paired Donation Network (PDN), which has ceased operations. In addition, there are several smaller private kidney exchanges in the US. They typically only involve one or a couple of transplant centers. These include an exchange at Johns Hopkins University and a single-center exchange at the Methodist Specialty and Transplant Hospital in San Antonio. Furthermore, there are now established kidney exchanges in the Netherlands, Canada, and England, as well as nascent ones in Portugal and Israel.

Kidney exchanges started with just using 2-cycles before also allowing 3-cycles and altruist-initiated chains (Roth et al. 2006). Since 2006, kidney exchanges have also incorporated never-ending chains, where the last donor in a chain serves as an altruist in a later match run to initiate a new chain (Rees et al. 2009). This approach is now included at least in the three leading kidney exchanges (UNOS, NKR, and APD).

Fielded kidney exchanges perform batch matching. The objective in the batch optimization engine is not to simply maximize the number of matches, but a weighted sum of the matches, where the weights—aka. priority points—are

decided by committees of medical personnel, computer scientists, and economists. At UNOS the current prioritization scheme was largely inherited from the US deceased-donor waiting list (which UNOS also runs) prioritization policy. The UNOS priority points take into account the following factors: do the donor and patient have zero antigen mismatch in tissue type, sensitization of the patient, prior organ donor status, pediatric status, wait time, geographic proximity, and other antibody specificities.

Appendix C: Additional Details about MAXCARD and MAXCARD-FAIR

In the main paper, we formally define and derive the MAXLIFE objective function. We now formally address the MAXCARD and MAXCARD-FAIR objectives. The MAXCARD-FAIR objective can be viewed as a generalized form of MAXCARD (that is, MAXCARD is just MAXCARD-FAIR with an empty set of vertices who are preferred by the objective).

Deciding which class of vertices are preferred is a complex ethical and medical decision. We use two common preference criteria in this paper: pediatric status and sensitization. Children (in the US, those who are under age 18) are typically treated preferentially in medical systems; we follow that rule here. For kidney exchange it has explicitly been articulated that such pediatric patients should be preferred not only because they have a lot of life left (barring their kidney disease) but also because having poor kidney function stunts growth. Some patients are *highly sensitized*, which means they are extremely unlikely to be medically compatible with a random organ. For these patients, finding a kidney is difficult (UNOS). The percentage of highly-sensitized patients in fielded kidney exchanges is high; over 60% of the patients in the UNOS kidney exchange are highly sensitized.

Figure 4 shows the evolution of the UNOS nationwide pool since inception, with each bar representing the pool at the time of a match run. Red bars show the portion of the pool that is marginalized (that is, either the patient is under the age of 18 or highly sensitized). UNOS currently prioritizes such patients significantly in its matching algorithm.

While defining fairness is a contentious issue in social science, a recent paper by Dickerson, Procaccia, and Sandholm (2014) formalizes two natural "fair" utility functions and shows how to optimize either of these functions in the deterministic or failure-aware *static* models of kidney exchange. We adapt the *weighted* fairness rule from that paper to FUTUREMATCH. The weighted fairness rule adjusts edge weights by some re-weighting function $\Delta : E \to \mathbb{R}^+$. A simple example re-weighting function is multiplicative:

$$\Delta^\beta(e) = \begin{cases} (1+\beta)w_e & \text{if } e \text{ ends in } V_P \\ w_e & \text{otherwise} \end{cases}$$

Here, $V_P \subseteq V$ is the set of vertices with preferred, marginalized patients. Intuitively, for some $\beta > 0$, this function scales the weight of edges ending in marginalized vertices by $(1+\beta)$. For example, if $\beta = 1.5$, then the optimization algorithm will value edges that result in a marginalized patient receiving a transplant at 250% of their initial weight (possibly scaled by factors like edge failure probability or chain position, as we discussed in the experimental section of the main paper).

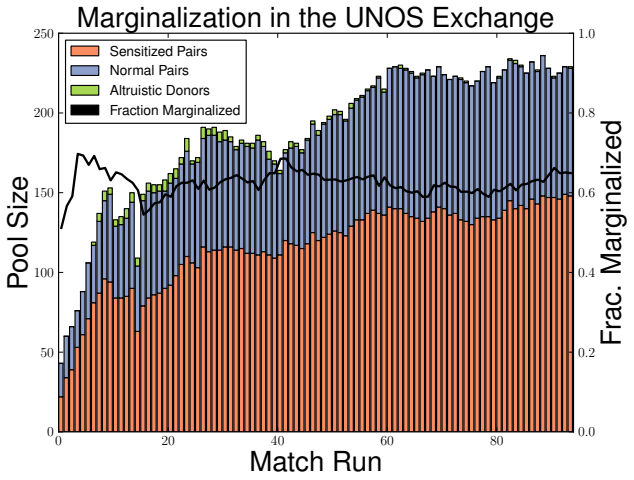

Figure 4: Evolution of the UNOS national kidney exchange. For each of 94 match runs (x-axis), the number of highly-sensitized or underage patients, non-highly-sensitized patients, and altruists are plotted (left y-axis), as well as the percentage of patients who are highly sensitized or underage as a percentage of the pool size (right y-axis).

For any $M \in \mathcal{M}$, let M' be the matching such that every edge $e \in E$ has augmented weight $\Delta^\beta(e)$. Then the MAXCARD-FAIR utility function u_Δ is defined in terms of the utilitarian MAXCARD utility function u applied to the augmented matching M', such that $u_\Delta(M) = u(M')$. In the experiments in the main paper, we varied the parameter β to empirically quantify its effects on each of the three objective functions.

Appendix D: Small Example of Potentials in Kidney Exchange

We now provide an example of setting *vertex potentials* in a reduced model of kidney exchange; note that, as described in our paper, we use a much richer set of features in our experiments. We must first select a set Θ of features representing different element types in the pool. Then, for each element type θ, assign some value $P_\theta \in \mathbb{R}$ that represents the expected potential usefulness of that kind of element to the pool over time. As an overly simplified pedagogical example, let $\Theta_{\text{ALT}} = \{\text{ALT}, \text{PAIR}\}$; potentials are assigned based on whether or not a vertex is an altruist or a patient-donor pair. Intuitively, $P_{\text{ALT}} \geq P_{\text{PAIR}}$; altruistic donors tend to be (much) easier to match because no returning edge is required to "close the cycle" at the end of a chain. These potentials are then subtracted out for each element in the objective.

Figure 5 shows a two time period example under Θ_{ALT}. Vertices a, p_1, and p_2 arrive in the first time period, while vertex p_3 arrives in the second time period. Assigning a (large enough) potential P_{ALT} results in the chains $\langle a, p_1 \rangle$ and $\langle a, p_1, p_2 \rangle$ having negative weight and thus not being matched in the first time period. However, when p_3 arrives, the chain $\langle a, p_1, p_2, p_3 \rangle$ may now have positive value (i.e., the utility of matching three pairs outweighs the learned potential of holding a back for another round) and can be

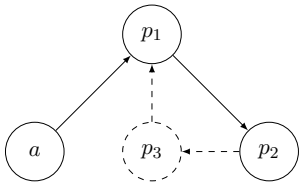

Figure 5: Example of potentials Θ_{ALT}. Pair p_3 appears in the second period. Myopic matching uses a to match two pairs; assigning a positive potential results in all three pairs matched without using a.

matched, or the 3-cycle $\langle p_1, p_2, p_3 \rangle$ has higher positive value and is matched instead, continuing to save the altruistic donor a for a longer chain in the future.

Appendix E: Dynamic Simulator and Verification

When learning potentials offline, it is important to mimic closely the behavior of the fielded exchange online. If the distribution of incoming potential types is different than expected, so too will be the estimates of potentials. We built a dynamic simulator of kidney exchange using data from the UNOS exchange (and APD (Ashlagi et al. 2011)). This work significantly extends that of Dickerson, Procaccia, and Sandholm (2013), which defined and experimentally evaluated a model of the evolution of dynamic kidney exchange. Critically, they did not perform dynamic optimization in that model—just myopic optimization applied sequentially in a dynamic model. They also sampled from a basic generator that is no longer accepted in the kidney exchange community (Saidman et al. 2006), while we sample from an accurate distribution—the historical UNOS exchange pool!

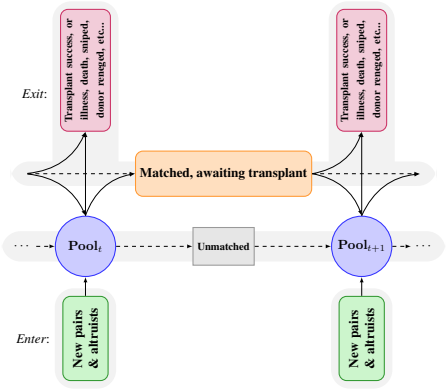

Figure 6: Dynamic kidney exchange.

Figure 6 portrays two time steps in our dynamic kidney exchange simulator. New pairs and altruistic donors enter the pool at each time period, while some leave the pool due to a variety of non-exchange-related reasons (becoming too ill to transplant, death, having a donor back out, finding a donor elsewhere). A matching is performed at each time period, which results in a set of matched pairs leaving the pool for $t > 0$ time periods. This reflects the length of time required to medically and logistically verify the implementability of the planned match. Matched patients then either leave the pool permanently after successfully receiving a kidney, or return to the pool after failing to receive a kidney. We set the relevant entrance and exit probabilities based on the real UNOS kidney exchange data (Kidney Paired Donation Work Group 2013).

A matching is determined at each time period based on either a *deterministic* or *failure-aware* clearing algorithm, which we briefly describe here. Both models compute an optimal matching $M^* = \arg\max_{M \in \mathcal{M}} u(M)$, where $u(M) = \sum_{c \in M} u(c)$. Here, $u(c)$ represents the utility of a cycle or chain c. In the deterministic model, $u(c) = \sum_{e \in c} f_w(e)$: that is, the sum of the weights of the constituent edges in a cycle or chain subject to the weight function w and potential mapping f learned in the main paper.

The deterministic model is susceptible to edge failures. For example, if a single edge in a 3-cycle fails, that entire cycle fails to execute. Similarly, if the third edge in a long chain fails, then the tail of that chain (after the failed edge) is cut off. Because of this dependence on other members of a cycle or chain, edge failures cannot simply be encoded via the edge weighting function w, which treats all edges (and their failures) independently. Let $q_e \in [0, 1]$ be the probability that an edge succeeds. Then the failure-aware model defines the *discounted* utility for a cycle c as $u(c) = \left[\sum_{e \in c} f_w(e)\right] \cdot \left[\prod_{e \in c} q_e\right]$, and the discounted utility for a chain $c = \langle e_0, e_1, \ldots, e_{k-1} \rangle$ as

$$u(c) = \left[\sum_{i=1}^{k-1}(1 - q_i)\sum_{j=0}^{i-1} f_w(e_j)\prod_{j=0}^{i-1} q_j\right] + \left[\sum_{i=0}^{k-1} f_w(e_i)\prod_{i=0}^{k-1} q_i\right].$$

To compute the optimal matching M^* at each time period, we use an adapted version of the standard integer programming-based batch clearing algorithm (Abraham, Blum, and Sandholm 2007; Dickerson, Procaccia, and Sandholm 2013; 2014) as a module in our dynamic simulator. In the main paper, we explored the effects of optimization under both models using each of the three objective functions we defined.

Appendix F: Experimental Method & Additional Experimental Results

Compatibility graphs are sampled with replacement from the set of all altruistic donors and patient-donor pairs who have ever participated (either successfully or unsuccessfully) in the UNOS exchange between Oct. 2010 and Jan. 2014, over a total of 94 match runs. Edges are drawn between two vertices in the graph if they pass the UNOS feasibility test, which determines compatibility with respect to patient and donor blood type, which kidney (left or right) is available, Hepatitis B/cytomegalovirus (CMV)/Epstein-Barr virus (EBV) positivity, creatinine clearance, blood pressure limits for the patient and donor, Body Mass Index (BMI) preferences, minimum and maximum age requirements for the donor, whether the donor and patient are willing to travel or accept a shipped organ, and other per-patient and per-donor requirements. The patient and donor features are used during the learning of the weight function for MAXCARD-FAIR (to determine marginalization status) and MAXLIFE

(to determine relative risk), and during the learning of potentials $\mathcal{P}_{\text{ABO}}$ (to determine blood types). Which exact features are used, and how, was discussed in the respective subsections earlier in the paper.

We draw edge failure probabilities in accordance with those published in the medical literature (Ashlagi et al. 2011). Critically, according to the experience of the exchange (APD) in that work, sensitization plays a large role in the probability of a match failing, with higher sensitization correlating with higher failure probability. These failure probabilities are incorporated directly into the optimization under the failure-aware model, and are incorporated directly into the *simulation* under *both* models (deterministic and failure-aware). A matched patient-donor pair leaves the pool for $t = 8$ weeks before receiving a confirmation of transplant success (and thus leaving the pool) or match failure (and thus reentering the pool, if the pair did not leave for other reasons like death or finding a donor elsewhere). Matching is performed once per week as is the current practice in the UNOS exchange; the total simulation occurs over $T = 24$ weeks. Each data point is based on at least 140 runs on separate generated realistic graphs.

We compare against the deterministic myopic matching algorithm, variants of which are employed by all fielded kidney exchanges. This matching algorithm does not explicitly take edge failure nor learned potentials into account during optimization; as described earlier, it finds a maximum weight matching (i.e., for each chain or cycle c, $u(c) = \sum_{e \in c} w_e$). We adapt this matching algorithm as in Dickerson, Procaccia, and Sandholm (2014) when appropriate to include equity constraints, using the re-weighting function Δ^{β} described earlier.

Additional Experimental Results

Figure 7 shows the increase in number of marginalized transplants fielded as the β fairness factor is increased. It also shows the relative gain in the MAXLIFE objective compared to myopic matching. From Table 2 in the main paper, MAXLIFE yielded roughly the same number of transplants, just weighted slightly toward those with longer expected graft survival. This correlation aligns with the general notion that easier-to-match patients also tend to be healthier overall, and thus similar vertices will tend to be favored (or disfavored) by both the MAXCARD and MAXLIFE objectives.

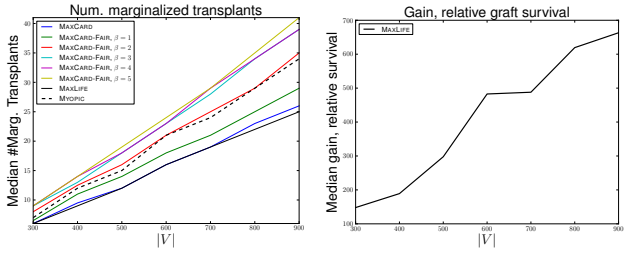

Figure 7: Median expected total number of marginalized transplants under FUTUREMATCH (left), and median expected gain in MAXLIFE (right).