MATCHING IN DYNAMIC IMBALANCED MARKETS

By

Itai Ashlagi, Afshin Nikzad, and Philipp Strack

COWLES FOUNDATION PAPER NO. 1852

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Challenge

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/mcha20

John Maynard Keynes Narrates the Great Depression: His Reports to the Philips Electronics Firm

Robert W. Dimand & Bradley W. Bateman

To cite this article: Robert W. Dimand & Bradley W. Bateman (26 Oct 2023): John Maynard Keynes Narrates the Great Depression: His Reports to the Philips Electronics Firm, Challenge, DOI: 10.1080/05775132.2023.2272544

To link to this article: https://doi.org/10.1080/05775132.2023.2272544

Published online: 26 Oct 2023.

CHALLENGE
https://doi.org/10.1080/05775132.2023.2272544

Routledge
Taylor & Francis Group

John Maynard Keynes Narrates the Great Depression: His Reports to the Philips Electronics Firm[†]

Robert W. Dimand and Bradley W. Bateman

ABSTRACT

In October 1929, the Dutch electronics firm Philips approached John Maynatd Keynes to write confidential reports on the state of the British and world economies, which he did from January 1930 to November 1934, at first monthly and then quarterly. These substantial reports (Keynes's November 1931 report was twelve typed pages) show Keynes narrating the Great Depression in real time, as the world went through the US slowdown after the Wall Street crash, the Credit-Anstalt collapse in Austria, the German banking crisis (summer 1931), Britain's departure from the gold exchange standard in August and September 1931, the US banking crisis leading to the Bank Holiday of March 1933, the London Economic Conference of 1933, and the coming of the New Deal. This series of reports has not been discussed in the literature, though the reports and surrounding correspondence are in the Chadwyck-Healey microfilm edition of the Keynes Papers. We examine Keynes's account of the unfolding events of the early 1930s, his insistence that the crisis would be more severe and long-lasting than most observers predicted, and his changing position on whether monetary policy would be sufficient to promote recovery and relate his reading of contemporary events to his theoretical development.

Introduction

On October 23, 1929, just as Wall Street began to crash[1] and the world economy moved into exceptionally interesting times, Dr. H. F. van Walsem, counsel and secretary to the Dutch electronics firm N. V. Philips Gloeilampenfabrieken[2], wrote to "J. M. Keynes, Esq., C.B. Cambridge" asking him to write a monthly letter to the firm's Economic Intelligence Service about the state of the British economy and the world economy. John Maynard Keynes's letters to Philips, monthly from January 1930 to November 1931 and then, because of budget cuts to Philips's Economic Intelligence Service, quarterly from February 1932 to November 1934, show Keynes narrating the events of the Great Depression as they occurred, and reveal his perception of the convulsions of the

CONTACT Robert W. Dimand rdimand@brocku.ca, robert.dimand@yale.edu Department of Economics, Brock University, St. Catharines, Ontario L2S 3A1, Canada.

[†]Unpublished writings of JOHN MAYNARD KEYNES copyright The Provost and Scholars of King's College Cambridge 2023.

(Fall 2023: Visiting Professor of Economics, Yale University)
Robert W. Dimand is a professor of economics at Brock University, Canada, and a visiting professor at Yale University, and author of *The Origins of the Keynesian* and an editor of *The Elgar Companion to John Maynard Keynes*.
Bradley W. Bateman is president emeritus of Randolph College, Virginia, coauthor of *Capitalist Revolutionary: John Maynard Keynes* and an editor of *The Cambridge Companion to Keynes*.

world economy as he wrote his *General Theory of Employment, Interest and Money* (1936). This substantial body of Keynes's commentary on economic fluctuations (the November 1931 letter alone is twelve typed, double-spaced pages) has hitherto been neglected in the literature on Keynes. Keynes's reports and the associated correspondence, preserved in the Keynes Papers at King's College, Cambridge, are included in the 1993 Chadwyck-Healey microfilm edition of the Keynes Papers (section BM/5 Memoranda Exchanged with Business Houses), but the expense of this edition (which was sold only as a complete set of 170 reels of microfilm, priced at £9,700 or $17,000, plus $175 for a hardcover catalogue, Cox 1993) meant that only a few copies were sold. According to the WorldCat catalogue, there are five sets in libraries in the United States (Library of Congress, Harvard, Yale, Ohio State, and University of Texas at El Paso), two in Great Britain (Universities of Oxford and Sheffield), one in Canada (Victoria University in the University of Toronto) and a few in Germany (Göttingen), Italy and elsewhere but surprisingly little use has been made even of these copies of Keynes's letters to N. V. Philips. Neither Moggridge (1992) nor Skidelsky (1983–2000, 2003), major biographies of Keynes by the authors who know the Keynes Papers best, mentions Keynes's reports to Philips (but Backhouse and Bateman 2011, 129, have a paragraph about Keynes's July 1930 report). As Jacqueline Cox (1995, 171) notes, the thirty volumes of Keynes's *Collected Writings* (1971–1989) include "only a third of the bulk classified as economic" in the Keynes Papers at King's and do not include Keynes's philosophical papers there, while "the personal papers were barely touched." Donald Moggridge (2006, 136–137) observes that "There has, inevitably, been heavier use of the Keynes Papers in King's College Cambridge, which have the advantage of being available elsewhere on microfilm, than, say, his papers in the National Archives or his correspondence with his publishers, the last of which reveals the risks of depending on the Cambridge collection alone." A vast amount of research has been done about Keynes and his economics, yet not all the relevant material has been explored (see Backhouse and Bateman 2006, Dimand and Hagemann 2019).

These reports reveal Keynes's reading of what was happening in the British and world economies through the first four years of the Great Depression, and provide the empirical counterpart to the record of Keynes's theoretical development in this period given by notes taken by students at Keynes's lectures from 1932 to 1935 (Rymes 1987, 1989, Dimand 1988, Dimand and Hagemann 2019). After the success of *The Economic Consequences of the Peace* (1919), Keynes no longer needed to be paid for lecturing, and so gave a single series of eight lectures each year, on the subject of whatever book he was writing at the time, so his lectures from 1932 to 1935 are in effect annual drafts of the book that became *The General Theory*. These lectures at Cambridge and the reports to N. V. Philips on what was happening in the economy provide theoretical and empirical supplements to Keynes's *Collected Writings* (1971–1989), respectively, in following Keynes's intellectual development in the Great Depression, from *A Treatise on Money* (1930) to *The General Theory* (1936). In Keynes's workload, his reports to Philips from 1930 to 1934 took the place of the London and Cambridge Economics Service Special Memoranda on commodity markets that he wrote from 1923 to 1930 (Keynes [1923–30] 1983, 267–647), which provided an empirical counterpart to his normal backwardation theory of futures contracts ([1923] 1983, 1930, Chapter 29).

Replying on October 31 to von Walsem's letter inviting him to write the monthly letter to the firm's Economic Intelligence Service, Keynes was "quite ready to discuss this proposal with one of your representatives" but wished to clarify "that there will be no question of the publication of the letters and that they will be purely for the information of your own people" – and that "it would not be practicable to me to undertake such work except in return for a somewhat substantial fee which might be higher than you would be willing to offer." On November 4, von Walsem assured him that the letters would not be published and "There are only two persons who, though not in our service, are closely related to our firm, who also receive a copy of our Intelligence Service which they, however, are bound to consider as absolutely confidential." He suggested £100 a year. On November 13, Keynes, having "considered your kind proposal in relation to the fees which I have received on previous occasions for somewhat analogous work," offered to undertake the task for an initial six months, for £150 a year[3]. Although Van Walsem had initially asked for the suggestion of other authors if Keynes preferred not take on the task at the suggested £100 a year, and Keynes equally pointedly offered to suggest such alternative authors if Philips did not care to pay £150 a year, Van Walsem accepted Keynes's terms for Philips on November 22: "We think it desirable that one of our gentlemen will see you in order to discuss some details in the first half of December next."

In the event two representatives of Philips (Messrs. Sannes and du Pré) met with Keynes for a discussion summarized "for good order's sake" by van Walsem on December 21, 1929 (by which time van Walsem had already received a December 18 note by Keynes on the Australian exchange position). He recorded agreement that Keynes's monthly letter would treat "some important factor in the development of the British economic situation and give your opinion as to its effects on trade in general and on our business in particular. Also you will draw our attention to important events in the domains especially interesting us, in so far as these come to your knowledge … Whenever you think it necessary you will give us your views on the situation in different parts of the British Empire or eventually of other countries. If possible we shall suggest [to] you special points to be considered in your letters." Von Walsem wrote again on June 21, 1930 to confirm "that the arrangement has given us full satisfaction so that we are willing to continue on the same terms" and enclosed a cheque for 75 pounds. The arrangement also satisfied Keynes; he wrote on January 1, 1931, that "I have enjoyed preparing the letters." Keynes's letters balanced opinions about trade in general with observations about matters affecting Philips more specifically. Thus on January 11, 1930, Keynes stated that "The Factory capacity for Radio Sets seems to have become quite appalling during 1929" before proceeding more generally "to take this opportunity of emphasizing the anxiety which is felt here about the Australian position … I think that Australia may have more difficulties with her balance of trade during the coming year than the Argentine."[4]

The Slump of 1930: Investment, Debts and Deflation

Keynes's April 1930 letter suggested that, although a general improvement had not yet arrived, "there are a fair number of indications that we may be somewhere in the neighborhood of the bottom point." In particular, "the continuance of cheap money, and even more the expectation of such continuance, is bound to be effective in the situation in the course of a few months," but the effect on employment would be slower than on business feeling and the Stock Exchange and "it would not be surprising to see British unemployment figures go on mounting even to the neighborhood of 2,000,000 up to the end of this calendar year. … The effect of many rationalization schemes now in train will be for some time to come to improve profits rather than employment." With a large amount of Australian gold en route to the Bank of England, "there is less anxiety about the British exchange position than there has been for a very considerable time past" and Keynes expected the creation of the Bank for International Settlements to have a positive effect on confidence, a foreshadowing of his emphasis at Bretton Woods on the importance of designing appropriate international monetary institutions. Keynes doubted that the Federal Reserve Board would reverse its cheap money policy "until business and employment in the United States is a great deal better than it is now." This emphasis on expectations would be characteristic of Keynes's *General Theory* (although equally in line with Irving Fisher's quantity theoretic concern with expected inflation), as is the measurement of the ease of monetary policy by the cheapness of money, that is, by low nominal interest rates. Because nominal interest rates (especially short-term rates such as the Treasury Bill rate) were very low in a period of deflation, the Federal Reserve Board continued to view monetary conditions as easy throughout what Milton Friedman and Anna Schwartz (1963) later termed the "Great Contraction" of the US money supply (during which the monetary base increased, but not by enough to offset the rise in currency/deposit and reserve/deposit ratios), despite Fisher drawing the attention of his former student, Federal Reserve Governor Eugene Meyer, to the statistics on the shrinkage of the money supply, the sum of currency and demand deposits (Cargill 1992, Dimand 2019).

On June 24, 1930, H. du Pré emphasized that, "In reply to your remarks about the character of your monthly letters, we assure you that we leave it entirely to you to judge in each case which are the topics which are most worth being discussed by you." Nonetheless, "There is one question upon which we particularly should like to have your opinion." Keynes's monthly letters had repeatedly stated that recovery depended on the bond market becoming more active, with new loans being used not just for the refunding of floating debt but for new productive investment. "But on the other hand these last months many articles in the economic press" saw excessive capacity in many industries; "in other words that the world has first to grow into a productive apparatus which is too big for immediate needs. If this should be true, can a renewed investment-activity soon be hoped for, and if it soon comes, would it really do good? Of course there would be less unemployment in a number of industries; but would not prices of consumptive commodities, and so cost of living, rise? And especially it might turn out after some time, that the new activity has only added to the – supposed – actual over-investment, so that the disequilibrium would only be greater. It may of course be that entirely new industries are going to take the lead, but we do not yet see any that are very likely to do so. We should be much obliged if you would solve this puzzle for us or at least give your views on the pretended overcapacity and its probable effects on future developments in your next letter." This letter sheds light on the audience for Keynes's reports in the secretariat of N. V. Philips: not just salesmen looking for tips about the market for radio sets in Great Britain or elsewhere, but thoughtful businessmen pondering sophisticated economic issues such as the dual nature of productive investment in creating demand while increasing capacity (a problem to which the warranted growth rate of Harrod 1939 was an attempted solution).

In his July 1930 letter (seven typed pages, plus a six-page note on the bond market), Keynes warned that "it is now fully clear the world is in the middle of an international cyclical depression of unusual severity … a depression and a crisis of major dimensions … I believe that the prevailing opinion in the United States is still not pessimistic enough and is relying too much on a recovery in the early autumn, an event which is, in my opinion, most improbable. Nothing is more difficult than to predict the date of recovery. But all previous experience would show that a depression on this scale is not something from which the recovery comes suddenly or quickly." He felt that "The optimism of Wall Street and the hoarding tendencies of France may prevent any real recovery of the International Loan Market this year" and considered whether this might lead to "a psychological atmosphere in which really drastic scientific measures will be taken by Great Britain and the United States in conjunction to do what is humanly possible to cause a turn of the tide next spring. But one is traveling here into the realm of the altogether uncertain and unpredictable." In contrast, the Harvard Economic Society (founded by Harvard economics professors Charles J. Bullock and Warren Persons) stated in its weekly letter on June 28, 1930, that "irregular and conflicting movements of business should soon give way to sustained recovery" and on July 19 that "untoward elements have operated to delay recovery but the evidence nevertheless points to substantial improvement" (quoted by Galbraith 1961, 150, see also Walter Friedman 2014).

Responding to du Pré's query, Keynes reiterated that recovery would be preceded by "a substantial fall in the long-period rate of interest … leading in due course to the recovery of investment." But now he explained that he was not thinking of investment in manufacturing industry, "the world's capacity for which is probably quite ample for the present." Even at the highest estimate, the total cost of bringing Britain's industrial plant up to date "would not use up the country's savings for more than, say, three months. Moreover, when expected profits are satisfactory the rate of expenditure by manufacturing industry in fixed plant is not very sensitive to the rate of interest."

"On the other hand," in contrast to manufacturing, "the borrowing requirements for building, transport and public utilities are not only on a far greater scale, but are decidedly sensitive to the rate of interest. If I were to put my finger on the prime trouble to-day, I should call attention to the very high rate of interest for long-term borrowers … the long-term rate of interest is higher to-day than it has been in time of peace for a very long time past. When, at the same time, there is a big business depression and prices are falling, it is not surprising that new enterprise is kept back at the present level of interest." He drew attention to "those who might be called distress borrowers, that is say countries which have an urgent need for borrowing to pay off existing debts, and are consequently ready to pay a very high rate of interest," citing prospective Austrian, Hungarian and Australian loans on the London bond market, and remarked that "the effect of the German Loan has been to supply the French Treasury with funds, which it has withdrawn from the French market and is keeping unemployed in the

Bank of France." Keynes's July 1930 letter (discussed briefly by Backhouse and Bateman 2011, 129) illuminates both his analysis of the present situation and the role of investment in his economics. His distinction between investment in manufacturing, responsive to expected profit rather than interest rates, and interest-sensitive investment in construction, transport and public utilities clarifies his theory of investment. Increased investment was crucial for recovery of the world economy, and low long-term interest rates were necessary for high levels of investment in construction, transport and public utilities, the largest part of investment (even if manufacturing investment depended more on expected profits). In regard to the current situation, Keynes explained the forces getting long-term interest rates high even when prices were falling and short-term interest rates were low, but felt that "progress has been made toward getting the necessitous borrowers out of the way." On the immediate practical level, Keynes's distinction between the determinants of the two categories of investment dealt with du Pré's question of how low long-term interest rates could stimulate investment given excess productive capacity in manufacturing. And yet, unlike Harrod (1939), Keynes's July 1930 letter did not come to grips with the theoretical point raised by du Pré, the dual character of investment in creating both demand and productive capacity.

Keynes's August 1930 letter dissented from the view widely held in the United States "even in responsible quarters, that we may expect an autumn recovery with some confidence … a good deal of the American optimism is based on analogies drawn from the date of recovery after the 1920-21 slump" (compare the Harvard Economic Society's statement on August 30 that "the present depression has about spent its force," quoted by Galbraith 1961, 150). He argued that "Too much emphasis cannot be laid on the really catastrophic character of the price falls of some of the principal raw materials since a year ago" (even larger than appeared from published index numbers, because those included a number of commodities subject to price controls), which "must profoundly affect the purchasing power of all overseas markets." Long-term interest rates remained high, reducing new capital investment. In contrast, Keynes considered general opinion about the British position to be "perhaps a little too pessimistic." Britain was already in a difficult position before the slump of 1929 and 1930, because of the 1925 return to the gold exchange standard at the prewar parity (over the eloquent protests of Keynes 1925). But the heavy unemployment in the slump was limited to textiles and heavy industry (iron and steel, coal, and shipbuilding), export-based sectors already hit by the return to gold at an overvalued exchange rate (in his December 1930 letter, Keynes stated that if textiles, iron and steel, and coal were omitted, there was practically no decline in the Index of Production from a year before and an improvement from two years before). Keynes explained that British unemployment statistics, when used in international comparisons, "probably overstate the case" since the British statistics included "a great many workers in definite employment, but working short time … It is even the case that workers taking their normal summer holidays are now included in the figures of the unemployed." According to *The Economist*, the aggregate profits of all British joint stock companies reporting their earnings in the first half of 1930 "were not only greater than in the previous year, but were larger than in any previous year. This was partly due to the prosperity of British Oil Companies operating abroad, but by no means wholly." Nor did Keynes share the worries of financial opinion in London (and so some extent his own previous letter to Philips) about "the constant dribble of gold to France."

In Keynes's September 1930 letter to Philips, he was "still of the opinion that real recovery is a long way off. But at the same time it seems to me not unlikely that we are at, or near, the lowest point … It is time, therefore, to cease to be a 'bear', even if it is not yet time to be a 'bull'." His February 1931 letter began, "Glancing through the letters of previous months, I find that they were all extremely pessimistic (with a brief lapse into modified optimism in September, corrected in October). Nevertheless, in the light of the actual course of events they were scarcely pessimistic enough. Nor do I see any reason for expecting any appreciable alleviation in the coming months." His September 1930 letter reported that "An extraordinary example of the way in which a situation can suddenly turn round, when a tendency has been greatly overdone, has been seen on the London Stock Exchange in the last two weeks. There has been no recovery of business in Great Britain to account for it. The real facts are much as they were a month ago. But market pessimism, aided by bear operations, had brought security prices down to an absurdly low level not justified by the circumstances … everyone knew in his heart that prices were falling to foolish levels. The result was that within a few days the prices of many leading securities had risen from 10 to 20 per cent." The stock market had diverged from any level that could be construed as reflecting underlying fundamentals, but then abruptly bounced back. Keynes again stressed that Britain was not doing as badly as the United States in the slump: the fall in the British index of production from the previous year "is certainly less than 10 per cent" whereas the US index of industrial production for July 1930 was 37% below that for July 1929.

Keynes's 1930 "October Letter" warned that, "The catastrophic increase in the value of money has raised the burden of indebtedness of many countries beyond what they can bear … in many parts of the world the fall of prices has now reached a point where it is straining the social system at its foundations. Agriculturists and other producers of primary materials are being threatened with ruin and bankruptcy all over the world. It is useless to expect a recovery of markets in such conditions" (and in his February 1931 letter he again warned that "The prospect of a long series of defaults [by debtor countries exporting raw materials] during 1931 is not be excluded"). All of the gains that Germany had received in the Young Plan for reparations compared to the Dawes Plan were obliterated because "the clause in the Dawes Plan by which her [Germany's] liabilities in terms of gold were to be modified in the event of a change in prices was not included in the Young Plan." Keynes declared himself "rather more pessimistic … than a month ago." He remarked that in Britain, "Very slight steps have been taken, as yet, in the direction of reducing wages, which is probably inevitable, but will not get anyone much further if all countries alike embark on wage-cutting policies."

These themes of Keynes's October 1930 letter to Philips, the danger of ruin and bankruptcy from price deflation in a world where debts are fixed in money terms and the futility of wage-cutting, appeared publically in his December article in *The Nation and Atheneum* on "The Great Slump of 1930" (reprinted in his *Essays in Persuasion*, 1931). There Keynes (1931, 138–139) warned that, since wage and price deflation increases the real burden of debt and wage cuts reduce purchasing power, "neither the restriction of output nor the reduction of wages serves in itself to restore equilibrium" and went on to emphasize that "Moreover, even if we were to succeed eventually in reestablishing output at the lower level of money-wages appropriate to (say) the pre-war level of prices, our troubles would not be at an end. For since 1914 an immense burden of bonded debt, both national and international, has been contracted, which is fixed in terms of money. Thus every fall of prices increases the value of the money in which it is fixed. For example, if we were to settle down to the pre-war level of prices, the British National Debt would be nearly 40% greater than it was in 1924 and double what it was in 1920; … the obligations of such debtor countries as those of South America and Australia would become insupportable without a reduction of their standard of life for the benefit of their creditors; agriculturalists and householders throughout the world, who have borrowed o mortgage, would find themselves the victims of their creditors. In such a situation it must be doubtful whether the necessary adjustments could be made in time to prevent a series of bankruptcies, defaults, and repudiations which would shake the capitalist order to its foundations" (see also Dimand 2011). Here, before Fisher (1932, 1933, see Dimand 2019), was the concern with the effect of deflation on the real value of nominal deflation that reappeared in Chapter 19, "Changes in Money Wages," of *The General Theory*, where Keynes (1936, 264) warned that "if the fall of wages and prices goes far, the embarrassment of those entrepreneurs who are heavily indebted may soon reach the point of insolvency – with severely adverse effects on investment."

Contested Budgets, Trade Balance and the Banking and Exchange Crises of 1931

In 1930, Keynes's "November Letter" argued that foreign opinion underestimated the financial strength that accompanied Britain's industrial weakness: "it is forgotten that the adverse tendencies of the foreign exchanges, until recently, have been due, not to the absence of a favorable foreign trade balance, but to the eagerness of British investors to take advantage of the high profits or high rates of interest obtainable abroad. In 1929 the British favorable balance available for new foreign investment was greater than that for any other country, greater even than that for the United States. The Bank of England's difficulties were due to the fact that the pressure of savers to take advantage of opportunities abroad was even greater." Subsequent events in Wall Street and elsewhere had made overseas investment less appealing to British savers, so that the Bank of England was holding twenty million pounds sterling more of gold than a year before. In his December 1930 letter, Keynes reported that, even though "The perpetual drain of gold to France provides a source of nervousness and irritation in the money market" and although thirty million pounds sterling of gold had moved from Britain to France in the previous three months, the Bank of England held twenty-two million pounds sterling more in gold than a year before (but Keynes's March 1931 letter reported that a drain of twenty million pounds sterling of gold from the Bank of England in the previous three months "causing nervous talk to prevail in London"). Despite Keynes's repeated insistence on the financial strength of sterling and the growing gold reserves of the Bank of England (less than a year before the crisis of August and September 1931 that forced Britain off the gold exchange standard), the underlying message was that capital mobility under fixed exchange rates would constrain even the Bank of England from trying to lower long-term interest rates to stimulate investment. Until Britain left the gold standard and allowed sterling to float, Keynes's letters to Philips monitored the strength of protectionist sentiment in the British Government, but he lost interest in tariff proposals once the exchange rate was no longer pegged (see Keynes 1931). But there was one bright spot for Britain: Keynes's February 1931 letter stressed that "It must not be overlooked that England is gaining enormously by the tremendous drop in the price of her imports as compared with that of her exports."

Keynes's April 1931 letter to Philips is notable for explaining that Britain's apparent budget deficit of £23.5 million for the fiscal year ending March 31 "is not as bad as it sounds, since this figure is reached after allowing for the repayment of £67,000,000 of debt. So that, apart from debt repayments, there was a surplus on the year's workings of £43,500,000. It must be doubtful whether any other country is showing so favorable a result. Even if the sum borrowed for the unemployment fund, which lies outside the budget[5], were to be deducted, there would still have on the year a net reduction of debt." The next year's was expected to be larger, but "If no debt were to be repaid, there would probably be no deficit, even for the forthcoming year." Keynes's May 1931 letter, reporting on the budget presented by Labor Chancellor of the Exchequer Phillip Snowden, noted that "there will still be some reduction of debt during the forthcoming year, though not on as large as a scale as formerly." A few months later, when Snowden and Prime Minister Ramsay MacDonald broke with their party to join the Conservatives in a National Government to deal with a budget and exchange crisis, Snowden found it convenient to overlook that the apparent budget deficit was an artifact of budgeting for a reduction in the national debt, and to denounce his former Labor Cabinet colleagues for endangering the savings of small depositors by having the Post Office Savings Bank lend to the Unemployment Insurance Fund, without mentioned that such loans were guaranteed by the Treasury or that he had neglected to inform his Cabinet colleagues of the borrowing (as Keynes indignantly explained in two paragraphs in the draft of his November 1931 letter, deleted from the final version).

Keynes's May 1931 letter is also notable, in light of the subsequent exchange crisis that forced Britain off gold in September, for insisting that "The improvement in the sterling exchanges and the better gold position of the Bank of England, as it appears in the public returns, are not deceptive and may be assessed at even more than their face value." He held that "When there is no longer serious pressure on the Bank of England's gold, the stage will be set for really cheap money throughout the world … It will not mean a recovery, but it will pave the way for the recovery of investment which must precede the recovery of prices and profits." Keynes again emphasized that "the fall in the prices of the commodities imported by Great Britain has been so much greater than the fall in the prices of her exports. On the visible trade balance Great Britain was £5,000,000 better off in the first quarter of 1931 than in either of the preceding years … Thus the main burden of the present crisis falls on the raw-material-producing countries, and Great Britain is likely to gain gold in spite of the immense decline of her exports."

By the next month, as the Credit-Anstalt collapsed in Vienna (see Schubert 1991), as French and American capital then took flight from Germany (see Balderston 1994), and as share prices slumped in London, Wall Street and on most European bourses, Keynes felt "that we are now entering the crisis, or panic, phase of the slump. I am inclined to think that we look back on this particular slump we shall feel that this phase has been reached in the summer months of 1931, rather than at any earlier date." He warned that "the consequences of a change in the value of money, as reflected in the prices of leading commodities, so violent as that which has occurred in the last eighteen months, cannot be regarded too gravely. Until prices show a material rise the whole fabric of economic society will be shaken. Each decline of commodity prices and each further collapse on the Stock Exchanges of the world brings a further group of individuals or institutions into a position where their assets doubtfully exceed their liabilities."

Looking across the Atlantic: The American Slump

Keynes's July 1931 letter focused on the United States, where 21% of the industrial population was unemployed with perhaps another 20% working only two or three days a week: "it is quite out of the question that there should be anything which could be called a true recovery of trade at any time within, say, the next nine months. The necessary foundations for such a recover simply do not exist." Many of the loans of small banks to farmers or secured by real estate "are non-liquid and probably impaired. Thus there is a strong desire for the utmost liquidity while obtainable on the part of the ordinary Bank; and general unwillingness to take any unnecessary risks or to embark on speculative enterprise, even where the risk may be actuarially a sound one. The nervousness on the part of the Bankers is accompanied by a nervousness of the part of their depositors … So there is quite a common tendency to withdraw money from the banks and keep resources hoarded in actual cash … It was estimated that in the country as a whole as much as $500,000,000 was hoarded in actual cash in this way" (see Fisher 1933, Friedman and Schwartz 1963, Bernanke 2000). Keynes stressed that, "The American financial structure is more able than the financial structure of the European countries to support the strain of so great a change in the value of money. The very great development of Bank deposit and of bondage indebtedness in the United States means that a money contract has been interposed between the real estate on the one hand and the ultimate owner of the wealth on the other. The depreciation in the money value of the real estate sufficient to cause margins to run off, necessarily tends therefore to threaten the solidity of the structure."

Keynes reported in his July 1931 letter that although US agricultural wages had fallen by 20 to 25%, and there had also been large cuts to wages in small-scale industrial enterprises, hourly wages were practically unchanged for two thirds of the workers in large-scale industrial enterprises while the hourly wages of the other third had been reduced by some 10%. In October 1934, however, Keynes stated in his Cambridge lectures that "Labor will and has accepted reductions in money wages, in the USA in 1932, and it will not serve to reduce unemployment" with one student's notes calling the money-wage reductions "catastrophic" (Rymes 1987, 131).

Germany Defaults, Britain Abandons the Gold Parity

Turning from the United States, Keynes remarked near the end of his July letter that, "At the moment of writing there are heavy gold drains from London; but I do not think that this need be regarded with any undue alarm," a judgment that proved too sanguine.

More presciently, he added “The real danger in the situation comes from the possibility of the declaration of a general moratorium in Germany and the collapse of the mark [Germany defaulted on July 15]. The repercussion of such events on the solvency of the banking and money market systems of the world would be most serious.” The next month, in his August 1931 newsletter (dated August 4), Keynes reported that “the bulk of the remaining short-term German debt is due to British and American banks and accepting houses; many accepting houses being landed with what are certainly frozen and may prove doubtful debts. Their own credit has suffered with the inevitable result, since they were the holders of large foreign balances, of a drain of gold from London … it would seem to be only ordinary prudence to act on the assumption that, while worse developments in Germany are doubtless possible, even apart from this the general underlying position is worse than the ordinary reader of newspapers believes it to be.” While “Great Britain is suffering from the temporary shock to confidence due to the difficulties of the accepting houses,”[6] the situation of the world economy as a whole was more serious: “We are certainly standing in the midst of the greatest economic crisis of the modern world. Important though the German developments have been I would emphasize that these have been essentially consequences of deeper causes which are affecting all countries alike … For there is no financial structure which can withstand the strain of so violent a disturbance of values.” A handwritten postscript at the end of the typed August 1931 letter warns Keynes’s readers “not to be encouraged even by the appearance of apparently good news. The world financial structure is shaken and is rotten in many directions. Patching arrangements will be attempted, but they will not do much good, and it would be a mistake to place reliance on them.” The next day, August 5, Keynes, writing to Prime Minister J. Ramsay MacDonald to urge rejection of the May Report, stated that “it is now virtually *certain* that we shall go off the existing parity at no distant date … when doubts, as to the prosperity of a currency, such as now exist about sterling, have come into existence, the game’s up” (Keynes 1971–1989, Vol. XX, 591–593; Skidelsky 2003, 446), but he did not say so in print or to Philips – and he rejected, on patriotic grounds, a suggestion by O. T. Falk that the Independent Investment Trust, of which Keynes and Falk were directors, should replace a dollar loan with a sterling loan, which Keynes condemned as “a frank bear speculation against sterling.” The Independent Investment Trust lost £40,000 by not switching its financing (Keynes 1971–1989, Vol. XX, 611–612; Moggridge 1992, 528–529; Skidelsky 2003, 447).

It was not only the world financial structure that was shaken; so was the Secretary Department of N. V. Philips. On August 6, 1931, H. du Pré wrote plaintively to Keynes, “Though we could hardly expect otherwise from your former letters, we note that you are not at all optimistic about the developments in the latter part of this year. These last weeks we read in the papers some statements from several Americans (among them people of authority), which hold a somewhat more cheerful view for the coming months. Must we infer from your letter that they are still, or again, too optimistic or is it possible that since your return from America[7] there have been some improvements, which may lead one to expect some improvement at least for the autumn?” Even Roger Babson, who had made his reputation by being bearish about the stock market in September 1929 (as he had been since 1926), was bullish by early 1931 (see W. Friedman 2014).

Keynes's reply on August 12 crushed any hopes: "In response to your enquiry, nothing has happened to make me more optimistic. As regards America, I consider that recovery this autumn is altogether out of the question. But the minds of all of us are of course dominated by the European and indeed the world situation. This still seems to me to be, as I have already described it, more serious than the general public know. I should recommend as complete inaction as is possible until further crises, or further striking events of some kind or another have occurred to clear up the situation."

Keynes's September letter (dated September 10, 1931), after the Conservative-dominated National Government displaced Labor, warned that "the hysterical concentration on Budgeting economy, which has also spread to the curtailment of expenditure by Local Authorities is calculated to produce unfavorable developments. For the widespread curtailment of expenditure is certain to reduce business profits and increase unemployment and lower the receipts of the Treasury, whilst it will do very little to tackle what is the fundamental problem, namely the improvement of the British Trade Balance. We seem likely to be faced by a period during which the balance of trade will not be sufficient to give confidence to foreign depositors."

It turned out, however, that one part of the cuts in government spending, the reduction in pay of the armed services, did indirectly dispose of the balance of payments problem. Since the government's version of equal sacrifice was that a vice-admiral earning £5 10s a day would lose 10 shillings a day (a reduction of 1/11), while naval lieutenants earning £1 7s a day and able-bodied seamen earning 5 shillings a day should each lose a shilling a day, reductions of 1/27 and 1/5, respectively (Muggeridge 1940, 109n), a naval mutiny erupted at Invergordon on September 16 (the first British naval mutiny since 1797), leading to abandonment of a fixed exchange rate on September 21 and a prompt 20% depreciation of sterling. Once the gold parity was abandoned, interest rates could be lowered without any balance of payments crisis. Commander Stephen King-Hall remarked "the strange combination of circumstances which caused the Royal Navy to be used by a far-seeing Providence as the unconscious means of … releasing the nation from the onerous terms of the contract of 1925 when the pound was restored to gold at pre-war parity … In 1805 the Navy saved the nation at Trafalgar; it may be that at Invergordon it achieved a like feat" (quoted by Muggeridge 1940, 111n). As for the budget deficit, Chancellor Snowden, who in the preceding Labor government had steadfastly blocked any reduction in the Sinking Fund contributions for paying down the national debt, now presented a budget reducing the annual Sinking Fund contribution by £20 million. Keynes declared in his October 1931 letter to Philips, "Great Britain's inevitable departure from the gold standard having occurred, it has been received with almost universal relief and in industrial circles a spirit of optimism is now abroad … Since the City and the Bank of England did their utmost to avoid the change, they feel that honor is satisfied. In other quarters the effect is to relieve a tension which was becoming almost unbearable … I have no doubt at all as to the reality of the stimulus which British business has obtained." Fisher (1935), assembling data on twenty-nine countries, found that recovery began only once a country abandoned the gold parity and was able to pursue a looser monetary policy (see Dimand 2003).

Keynes concluded his October 1931 letter, “The general passion for liquidity is bringing the value of cash in terms of everything else to so high a level as to be very near breaking point. This does not apply to Great Britain since her crisis was a balance of payments crisis rather than a banking crisis strictly so called. Thus the possibility of a general European and American banking crisis is the main risk, the possibility of which has now to be borne in mind.” The US banking crisis culminated in the “Bank Holiday” of March 1933, while all the major German and Italian banks passed into government ownership.

On November 3, 1931, Dr. du Pré was “very sorry to say that the necessity for the strictest economy which makes itself felt in all departments of our concern at present, impels us to an important curtailment of the budget of our Economic Intelligence Service” which would now issue bulletins every three months, instead of monthly. He asked Keynes for quarterly letters for £50 per annum, instead of monthly letters for £150 per annum. Keynes replied on November 9 that he read the letter “without any great surprise. I had been rather hesitating in my mind as to whether it is worth while to continue the arrangement on the new basis. But on the whole I feel that I should not like to break the friendly relations which have arisen between us, merely because times are bad.” He accepted the offer[8], asking to be reminded when each quarterly report was due, and enclosed his November letter stating that Britain was “to a considerable extent getting the best of both worlds since broadly speaking the countries from which we buy our food and raw materials have followed us off gold, whilst our manufacturing competitors have remained on the old gold parity.”[9] He felt that Continental observers were mistaken to think that Britain would want to return to gold: “Foreigners always underestimate the slow infiltration of what I have sometimes called ‘inside opinion’, whilst ‘outside opinion’ remains ostensibly unchanged. Then quite suddenly what ‘inside opinion’ becomes ‘outside opinion’. Foreigners are quite taken by surprise, but the change is really one which had been long prepared. In the later months of the old gold standard there was a hardly a soul in this country who really believed in it. But it was considered that it was our duty for fairly obvious reasons to do everything we possibly could to keep where we were.”

Keynes’s May 1932 quarterly letter stressed that, “The most important development, if one is thinking not so much of the moment but of laying the foundations for future improvement, is to be found in the return to cheap money, which was interrupted by the financial crisis of last summer and the departure from gold. I am more and more convinced in the belief, which I have held for some time, that an ultra-cheap money phase in the principal financial centers is an indispensable preliminary to recovery … Nevertheless it would be imprudent to expect too much at any early date from the stimulus of cheap money. The courage of enterprise is now so completely broken, that the effect on prices of money however cheap will be very slow. I consider it likely, therefore, that the cheap money phase may be extremely prolonged and that it may proceed to unprecedented lengths before it produces its effect.” He concluded, “For the time being the world is marking time, – waiting for it does not quite know what. I emphasize again the fact that the position in Great Britain, and in some of her Dominions, is relatively good. But for the time being, I see no light anywhere else … It would certainly be much too soon to take any steps whatever to be ready for a possible revival.”

Looking across the Atlantic: Hope from the New Deal

Keynes's August 1932 memorandum was notable for its explanation of why US stock prices had risen sharply and why that need not signal an end to the industrial crisis: the financial crisis had driven down stock prices until "the securities of many famous and successful companies were standing at little more than the equivalent of the net cash and liquid resources owned by those companies … the assets in question would either be worth nothing as a result of the general breakdown of contract, or must, in any circumstances apart from that, be worth a very great deal more than their quotations. Consequently, it is logical and right that the fear of their being worth nothing having been brought to an end, there should be a rapid recovery of the quotations on a very striking scale. It does not need a termination of the industrial crisis, or even an expectation of its early cessation, in order to justify the new levels."

In his February 1933 memorandum, commenting on the likely futility of the projected World Economic Conference, Keynes recalled that "I have myself put forward more drastic proposals for an international fiduciary currency, which would be the legal equivalent of gold. If this were agreed to, the position would be so much eased that various other desirable measures would also become practicable. I do not despair of converting British opinion to such a plan, but I am told that continental opinion would be almost unanimously opposed it." Keynes had contemplated such proposals long before Bretton Woods.

Keynes's August 1933 memorandum (actually mailed July 20, before Keynes left for holidays) held that "My own view is that President's Roosevelt's programme is to be taken most seriously as a means not only of American, but of world recovery. He will suffer set-backs and no one can predict the end of the story. But it does seem fairly safe to say that his drastic policies have had the result of turning the tide in the direction of better security not only in the United States, but elsewhere … Perhaps in the end President Roosevelt will devalue the dollar in terms of gold by 30 or 40 per cent." His November 1933 memorandum regretted "the failure of the President during his first six months to act inflation as well as talk it. In actual fact Governmental loan expenditure in the United States up to the end of September was on quite a trifling scale" but since then it seemed to be increasing: "if during the next six months the President is at last successful in putting into circulation a large volume of loan expenditure, I should expect a correspondingly rapid improvement in the industrial prosperity of America. This, if it occurs, would have a great influence on the rest of the world and especially on Great Britain … it might pave the way for a rate of improvement sufficiently rapid to deserve the name of real recovery." Keynes's February 1934 memorandum reported that in the United States "everything is moving strongly upwards. This is to be largely attributed to the fact that Governmental loan expenditure is now at last occurring on a large scale … the disbursement by the American Treasury of new money against borrowing has reached or is approaching $50,000,000 weekly and should maintain this rate for a few months to come." In his August 1934 memorandum, having visited the United States since his May memorandum, he found there "a recession which is somewhat more than seasonal," aggravated since his visit by a "failure of the corn crop … so acute as to be little short of a national disaster" but the actual and prospective level of US Government loan-financed expenditure made him optimistic about prospects for the US economy in the autumn and winter. He also reported that "the view is generally held in Great Britain that the gold block countries – including Holland not less than the others – cannot permanently maintain their present parity with gold without a disaster. Now or later it seems to us certain that the necessity for devaluation will be admitted." The reports end with Keynes's November 1934 memorandum, with no correspondence in the Keynes Papers concerning the end of his relationship with the Philips firm.

Conclusion: The Message of Keynes's Reports to Philips

Keynes's letters to the Philips electronics firm reveal he perceived events in the British and world economies from the beginning of 1930 through November 1934, and provide pungent and insightful commentary. These reports high-light the importance to Keynes of cheap money as a stimulus to investment – he was not just concerned with fiscal policy as the means to recovery, however much he placed emphasis from 1933 onward on the loan-financed expenditure of the Roosevelt Administration in the US. Keynes's response to a query from du Pré is particularly interesting about Keynes's distinction between those investment expenditures that are sensitive to interest rates and those that are not. The reports stress a theme discussed more briefly in Keynes's 1931 Harris Foundation lectures in Chicago (in Wright, ed., 1931) and in Chapter 19 of *The General Theory*, and at greater length by Irving Fisher (1932, 1933) (and later by Hyman Minsky 1975): since debt are contracted in nominal terms, a rise in the purchasing power of money increases the risk of bankruptcy, repudiation and default – and it is not just actual defaults that are costly, but also the perception of increased riskiness. Keynes recognized the exceptional seriousness of the Depression, dissenting firmly from predictions of an early recovery, and he saw clearly how defending overvalued gold parities forced central banks to keep interest rates high, instead of pursuing ultra-cheap money to restore investment. This hitherto-neglected body of evidence allows one to watch the unfolding of the world economic crisis of the early 1930s through Keynes's eyes, extraordinary events as viewed and narrated by an extraordinary economist. At £12 10s per report (by no means a trivial sum at the time), N. V. Philips certainly got their money's worth.

Notes

1. "Thursday, October 24, is the first of the days which history – such as it is on the subject – identifies with the panic of 1929" (Galbraith 1961, 103–104), but already on Monday, October 21, Irving Fisher had characterized the fall in stock prices as just the "shaking out of the lunatic fringe" and on Tuesday, Charles Mitchell of the National City Bank declared that "the decline has gone too far" (Galbraith 1961, 102).
2. Philips Incandescent Lamp Works, later Philips Electronics, successor to a firm founded by Lion Philips (originally Presburg), maternal uncle of Karl Marx (Gabriel 2011, 44, 110, 291-93, 295, 299, 315, 334, 366). Although relations between uncle and nephew were "strained by politics" (Gabriel 2011, 291), Mary Gabriel (2011, 299) refers to Marx's "fund of last resort, his uncle … He had sold himself to this pragmatic businessman as a successful writer only temporarily short of cash." Gabriel (2011, 642) remarks that "Marx's dabbling in the stock market has been questioned by some scholars, who believe he may simply have wanted his uncle to believe he was engaged in 'capital' transactions, not *Capital*." After the death of Lion

Philips, his sons did not reply to Marx's letter asking for help with his daughter Laura's wedding (Gabriel 2011, 364). Anthony Sampson (1968, 95) reported that the firm's chairman Frits Philips was "a keen Moral Rearmer and a fervent anti-communist, embarrassed by the fact that his grandfather was a cousin of Karl Marx."

3. For a sense of what £150 a year might have meant to Keynes: Moggridge (1992, 508, 585) and Skidelsky (2003, 417–418, 519, 565) report that Keynes's net worth fluctuated from £44,000 at the end of 1927 to £7,815 at the end of 1929, then rising to over £506,222 at the end of 1936, dropping again to £181,244 at the end of 1938. The offer from Philips came at a particularly low point in his finances. According to Skidelsky (2003, 265) "investment, directorship and consultancy income" accounted for more than 70% of Keynes's income between 1923-24 and 1928-29 (including £1,000 a year as chairman of National Mutual Life Assurance), books and articles for another 20%, leaving no more than a tenth of income from such academic sources as teaching, examining, being secretary of the Royal Economic Society and editor of its journal, and being Bursar and a Fellow of King's College.
4. However, writing to Keynes on January 21, H. du Pré was moved "to remark that the latest figures from the Argentine which, according to the handwritten note at the bottom of your letter, you intended to enclose, were not received here, so that we cannot give you an opinion about their importance for us."
5. When the majority report of the May Committee on National Expenditure projected on July 31, 1931, that the budget deficit for 1931-32 would be £120 million, necessitating £96 million of cuts to unemployment benefits, road construction, and government and armed forces pay, it counted all borrowing by the Unemployment and Road funds as "public expenditure on current account" as well as "the usual provision for the redemption of debt" of £50 million (Winch 1969, 126–130). Keynes accused the majority on the May Committee of not "having given a moment's thought to the possible repercussions of their programme, either on the volume of unemployment or on the receipts of taxation" – he estimated it would add 250,000 to 400,000 to the unemployed, and reduce tax receipts by £70 million (*New Statesman and Nation*, August 15, 1931; Keynes 1971-89, Vol. IX, 141–145; Winch 1969, 130, Skidelsky 2003, 446).
6. With regard to Britain, Keynes noted that "There is, however, tremendous pressure of public opinion towards the Government Economy, which means in the main a reduction in the salaries of Government employees and of the allowances of the unemployed. It is equally difficult for the present [Labour] Government either to refuse or concede concessions to this trend of opinion. But if a movement in this direction takes place, which is still most doubtful, it remains exceedingly open to argument whether the result on the actual level of unemployment will be favourable."
7. Keynes had given three Harris Foundation Lectures on "An Economic Analysis of Unemployment" at the University of Chicago in June and July 1931, published in Quincy Wright, ed. (1931), and reprinted in Keynes (1971-89), Vol. XIII. These lectures mostly expounded the analysis of Keynes's *Treatise*, but the third lecture also examined the debt-deflation process, the undermining of the financial structure by an increase in the real value of debts and fall in the nominal value of collateral (Keynes 1971-89, Vol. XIII, 359–361, see Dimand 2011).
8. He also raised a "small personal matter", asking for advice on buying a new wireless set that would "have a thoroughly good loud speaker, both for voice and music reproduction and should be able to pick up distant stations such as Moscow."
9. A passage crossed-out in the draft of Keynes's November 1931 letter, in the section discussing the general election, stated that, "As has been the case in the last three or four General Elections, it is that old wretch Lord Rothermere [publisher of the *Daily Mail*] who has been dead right. It is said that he has made a profit on the crisis of £100,000, buying majorities on the Stock Exchange." Skidelsky (2003, 472) relates that Keynes "consistently lost money (his own and his friends') on the results of general elections."

References

Backhouse, Roger E., and Bradley W. Bateman, eds. 2006. *The Cambridge Companion to Keynes*. Cambridge, UK: Cambridge University Press.

Backhouse, Roger E., and Bradley W. Bateman. 2011. *Capitalist Revolutionary: John Maynard Keynes*. Cambridge, MA: Harvard University Press.

Balderston, Theo. 1994. "The Banks and the Gold Standard in the German Financial Crisis of 1931." *Financial History Review* 1 (1):43–68. https://doi.org/10.1017/S0968565000001554

Bernanke, Ben S. 2000. *Essays on the Great Depression*. Princeton, NJ: Princeton University Press.

Cargill, Thomas F. 1992. "Miscellany: Irving Fisher Comments on Benjamin Strong and the Federal Reserve in the 1930s." *Journal of Political Economy* 100 (6):1273–7. https://doi.org/10.1086/261861

Chadwyck-Healey. 1993. *The John Maynard Keynes Papers in King's College, Cambridge, 170 Reels of Microfilm*. Cambridge, UK: Chadwyck-Healey Ltd.

Cox, Jacqueline. 1993. *A Catalogue of the John Maynard Keynes Papers in King's College, Cambridge*. Cambridge, UK: Chadwyck-Healey Ltd.

Cox, Jacqueline. 1995. "Keynes: An Archivist's View." In *New Perspectives on Keynes,* Annual Supplement to History of Political Economy. Vol. 27, edited by Allin F. Cottrell and Michael S. Lawlor, 163–75. Durham, NC: Duke University Press.

Dimand, Robert W. 1988. *The Origins of the Keynesian Revolution*. Aldershot, UK: Edward Elgar Publishing.

Dimand, Robert W. 2003. "Irving Fisher on the International Transmission of Booms and Depressions through Monetary Standards." *Journal of Money, Credit and Banking* 35 (1):49–90. https://doi.org/10.1353/mcb.2003.0002

Dimand, Robert W. 2011. "The Consequences to the Banks of the Collapse of Money Values,' 1931 and 2009." In *Perspectives on Keynesian Economics*, edited by Arie Arnon, Jimmy Weinblatt, and Warren Young, 233–45. Heidelberg: Springer Verlag.

Dimand, Robert W. 2019. *Irving Fisher*. London: Palgrave Macmillan, Great Thinkers in Economics.

Dimand, Robert W., and Harald Hagemann, eds. 2019. *The Elgar Companion to John Maynard Keynes*. Cheltenham, UK.

Fisher, Irving. 1932. *Booms and Depressions*. New York: Adelphi.

Fisher, Irving. 1933. "The Debt-Deflation Theory of Great Depressions." *Econometrica* 1 (4):337–57. https://doi.org/10.2307/1907327

Fisher, Irving. 1935. "Are Booms and Depressions Transmitted Internationally through Monetary Standards?" *Bulletin de L'Institut Internationale de Statistique* 28 (1):1–29. Reprinted in Dimand (2003).

Friedman, Milton, and Anna J. Schwartz. 1963. *A Monetary History of the United States, 1867–1960*. Princeton, NJ: Princeton University Press for the National Bureau of Economic Research.

Friedman, Walter A. 2014. *Fortune Tellers: The Story of America's First Economic Forecasters*. Princeton, NJ: Princeton University Press.

Gabriel, Mary. 2011. *Love and Capital: Karl and Jenny Marx and the Birth of a Revolution*. New York: Little, Brown.

Galbraith, John Kenneth. 1961. *The Great Crash 1929, with New Introduction*. Boston, MA: Houghton Mifflin.

Harrod, Roy F. 1939. "An Essay in Dynamic Theory." *The Economic Journal* 49 (193):14–33. https://doi.org/10.2307/2225181

Keynes, John Maynard. 1983 [1923]. "Some Aspects of Commodity Markets." *Manchester Guardian Commercial, Reconstruction in Europe* 29 March 1923; reprinted in Keynes (1971–89), XII:255–66.

Keynes, John Maynard. 1983 [1923–30]. *London and Cambridge Economic Service Special Memoranda*, as reprinted in Keynes (1971–89), XII: 267–647.

Keynes, John Maynard. 1919. *The Economic Consequences of the Peace*. London: Macmillan.

Keynes, John Maynard. 1925. *The Economic Consequences of Mr. Churchill*. London: Leonard and Virginia Woolf, at the Hogarth Press.

Keynes, John Maynard. 1930. *A Treatise on Money*, 2 Volumes. London: Macmillan.

Keynes, John Maynard. 1931. *Essays in Persuasion*. London: Macmillan.

Keynes, John Maynard. 1936. *The General Theory of Employment, Interest and Money*. London: Macmillan.

Keynes, John Maynard. 1971–89. *The Collected Writings of John Maynard Keynes*, 30 volumes, general editors Donald E. Moggridge and E. A. G. Robinson, volume editors Elizabeth S. Johnson and Donald E. Moggridge. London: Cambridge University Press, for the Royal Economic Society.

Minsky, Hyman P. 1975. *John Maynard Keynes*. New York: Columbia University Press.

Moggridge, Donald E. 1992. *Maynard Keynes: An Economist's Biography*. London: Routledge.

Moggridge, Donald E. 2006. "Keynes and His Correspondence." In *The Cambridge Companion to Keynes*, edited by Roger E. Backhouse and Bradley W. Bateman, 136–59. Cambridge, UK: Cambridge University Press.

Muggeridge, Malcolm. 1940. *The Thirties: 1930-40 in Great Britain*. London: Hamish Hamilton.

Rymes, Thomas K., ed. 1987. *Keynes's Lectures: Notes of Students*. Ottawa: Department of Economics, Carleton University.

Rymes, Thomas K., ed. 1989. *Keynes's Lectures, 1932–35: Notes of a Representative Student*. London: University of Michigan Press.

Sampson, Anthony. 1968. *The New Europeans*. London: Hodder and Stoughton.

Schubert, A. 1991. *The Credit-Anstalt Crisis of 1931*. Cambridge, UK: Cambridge University Press.

Skidelsky, Robert. 1983–2000. *John Maynard Keynes*, 3 volumes. London: Viking.

Skidelsky, Robert. 2003. *John Maynard Keynes, 1883-1946: Economist, Philosopher, Statesman*. London: Macmillan.

Winch, Donald. 1969. *Economics and Policy: A Historical Study*. New York: Walker and Company.

Wright, Quincy, ed. 1931. *Unemployment as a World Problem: The Harris Foundation Lectures*. Chicago, IL: University of Chicago Press.

Review of Economic Studies (2023) **90**, 1084–1124 https://doi.org/10.1093/restud/rdac044

Advance access publication 2 August 2022

Matching in Dynamic Imbalanced Markets

ITAI ASHLAGI
MS&E, Stanford
AFSHIN NIKZAD
Economics, University of Southern California
and
PHILIPP STRACK
Economics, Yale University

First version received April 2019*; Editorial decision September* 2021*; Accepted July* 2022 (*Eds.*)

We study dynamic matching in exchange markets with easy- and hard-to-match agents. A greedy policy, which attempts to match agents upon arrival, ignores the positive externality that waiting agents provide by facilitating future matchings. We prove that the trade-off between a "thicker" market and faster matching vanishes in large markets; the greedy policy leads to shorter waiting times and more agents matched than any other policy. We empirically confirm these findings in data from the National Kidney Registry. Greedy matching achieves as many transplants as commonly used policies (1.8% more than monthly batching) and shorter waiting times (16 days faster than monthly batching).

Key words: Matching, Kidney Exchange, Imbalance, Waiting times.

JEL Codes: C78, D47

1. INTRODUCTION

We study how to optimally match agents in a dynamic random exchange market. Faster matching of agents reduces waiting times, but at the same time makes the market thinner, leaving more agents without a compatible partner. This trade-off naturally arises for kidney exchange platforms that seek to form exchanges between patient–donor pairs, whose patient cannot receive the donor's organ.[1] Waiting to match may increase the number of patients receiving a kidney, but this comes at a cost: receiving a transplant earlier not only improves the quality of life for the patient but also leads to substantial savings in dialysis costs for society.[2] In the last decade, kidney exchange

[1] For some early work on kidney exchange in static pools and the importance of creating a thick marketplace, see Roth, Sönmez and Ünver (2004, 2007).

[2] The savings from a transplant over dialysis is estimated by over $270,000 per year over the first 5 years (Held, McCormick, Ojo and Roberts, 2016).

The editor in charge of this paper was Andrea Galeotti.

platforms in the US gradually moved from matching roughly every month to matching daily.[3] Practitioners are concerned that this behaviour, some of which is driven by competition between kidney exchanges, is harmful, especially for the most highly sensitized patients. In contrast, kidney exchange programs in Canada, Australia, and the Netherlands match periodically every 3 or 4 months (Ferrari, Weimar, Johnson, Lim and Tinckam, 2014).

This article analyses the trade-off between agents' waiting times and the percentage of matched agents in dynamic markets. We find that, maybe surprisingly, matching greedily minimizes the waiting time and simultaneously maximizes the chances to find a compatible partner for *all* agents in sufficiently large markets. We further quantify the inefficiency associated with other commonly used policies like monthly matching using data from the National Kidney Registry (NKR).

To analyse this question, we propose a stochastic compatibility model with easy-to-match and hard-to-match agents. Easy-to-match agents can match with each other with probability $q > 0$ and with hard-to-match agents with probability $p > 0$, whereas hard-to-match agents can match only with easy-to-match agents with probability $p > 0$. The main focus of our analysis is on the case where the majority of agents are hard-to-match, which is in line with kidney exchange pools. This compatibility model captures two empirical regularities of the patient–donor data from the NKR. First, as the market grows large, the fraction of patient–donor pairs that are matched in a maximal matching does not approach 1, which is a consequence of the imbalance between different pairs' blood types in kidney exchange (Saidman, Roth, Sönmez, Ünver and Delmonico, 2006; Roth *et al.*, 2007). Second, as the market grows large, the fraction of agents that cannot be matched in *any* matching goes to zero.[4] Our parsimonious two-type model captures the above regularities and no single-type model can account for both of them (Propositions 1 and 2).

We study a dynamic model based on the two-type compatibility structure in which easy- and hard-to-match agents arrive to the market according to independent Poisson processes with rates m_E and m_H. Agents depart exogenously at rate d. The market-maker observes the realized compatibilities and decides when to match compatible agents. We evaluate a policy based on two measures: *match rate* and *waiting time*. The match rate is the probability with which an agent is matched. The waiting time is the average difference between the time an agent arrives and the time she leaves, matched or unmatched. Our two-type model captures the potential trade-off between match rates and waiting times that concerns practitioners in an intuitive way: matching quickly could lead to easy-to-match agents being paired with each other thereby making it more difficult for hard-to-match agents to find a partner and thus potentially decreasing the overall match rate.

We start by analysing the *greedy policy*, which matches every agent upon its arrival if possible. We derive the distribution of the number of hard- and easy-to-match agents waiting in the market in steady state. As the market grows large, many hard-to-match agents will wait in the market for a compatible partner at any point in time. Consequently, almost every easy-to-match agent is matched with a hard-to-match agent immediately upon arrival and the probability that an easy-to-match agent leaves the market unmatched converges to zero (Proposition 3). As their match rate is close to one and their waiting time is close to zero, the greedy policy is asymptotically optimal for easy-to-match agents in large markets. As hard-to-match agents are incompatible with each other and almost every easy-to-match agent is matched with a hard-to-match agent, the greedy policy also maximizes the match rate of hard-to-match agents. Perhaps less intuitively, the greedy policy minimizes the waiting time of hard-to-match agents compared to *any* other policy

[3] The NKR and the Alliance for Paired Donation (APD) search for matches on a daily basis, whereas the United Network for Organ Sharing (UNOS) searches for matches twice per week.

[4] A patient–donor pair cannot be matched in any matching if it cannot form a (two-way) exchange with any other patient–donor pair due to biological compatibility.

when the market grows large (Proposition 4). This holds since the greedy policy matches weakly more hard-to-match agents than any other policy. Little's law implies that the average number of hard-to-match agents waiting in the market is proportional to their waiting time and thus that the greedy policy will perform weakly better than any other policy in a sufficiently large market.

The main challenge in the proof is analysing the steady-state distribution of a two-dimensional Markov chain which keeps track of the number of easy- and hard-to-match agents waiting in the market. Using the Lyapunov function method, we show that the stationary distribution is concentrated around the solution of a fixed-point equation that describes the average numbers of easy- and hard-to-match agents waiting in the pool. These concentration bounds allow us to compute the agents' match rate and waiting time.

Next, we quantify the inefficiency associated with batching policies, which are commonly used. A batching policy periodically (e.g. monthly) matches as many agents as possible. We derive the waiting time and match rate under batching policies in large markets. Batching less frequently decreases the match rate and increases the waiting time. Therefore, in a large market, greedy matching dominates any batching policy with a fixed batch length, as it leads to *strictly shorter* waiting times and *strictly higher* match rates. We also compare batching and greedy matching in finite markets, where batching may outperform greedy matching. We find that for parameters in line with our kidney exchange application batching policies need to match more frequently than daily to not be outperformed by greedy matching.

We also analyse the *patient policy* introduced by Akbarpour, Li and Gharan (2020). This policy assumes that agents' exogenous departure times are observable. It matches an agent at her departure time if possible, and otherwise the agent leaves the market unmatched. We show that the patient policy leads to the same match rate as the greedy policy when the market becomes large. In both policies almost all easy-to-match agents are matched almost upon arrival in a large market. Moreover, hard-to-match agents wait longer (in first-order stochastic dominance) under the patient policy compared to the greedy policy. Quantitatively, the waiting time of hard-to-match agents under the greedy policy equals the waiting time under the patient policy multiplied by $\left(1-\frac{m_E}{m_H}\right)$ where m_E and m_H are the arrival rates of easy- and hard-to-match agents. For example, when one-third of agents are easy-to-match ($2m_E = m_H$), hard-to-match agents will wait twice as long under the patient policy.

Finally, we test whether the large-market predictions of our model hold in data from the NKR. These data differ from our assumptions along two dimensions: first, because of blood and tissue types, it does not match our stylized two-type compatibility structure. Second, it is unclear that the market is sufficiently large for our results to apply because only a finite number of agents arrive every year (around 360/year). Nevertheless, the data confirm the predictions of our model (Section 5): As the market becomes large, the waiting times of patient–donor pairs who are "easier" to match approach 0, but the waiting times of "harder" to match pairs do not. Moreover, batching policies result in no improvement to the match rate and lead to longer waiting times relative to greedy matching (c.f. Table 1). Finally, waiting times under the greedy policy are significantly lower than under the patient policy. At the same time, we do not find significant differences between the match rates under greedy and patient policies (Table 1 and Figure 9).

As mentioned earlier, practitioners expressed concerns that kidney exchange platforms in the US are adopting greedy algorithms, arguably due to competition.[5] That is, matching greedily can potentially squeeze the liquidity generated by easy-to-match pairs in an inefficient manner.

[5] See Ashlagi, Bingaman, Burq, Manshadi, Gamarnik, Murphey, Roth, Melcher and Rees (2018) who write: "..competition among KPD programs to produce transplants may have incentivized programs to perform match runs at high frequency, which raises a major concern that such frequent matching may lead to fewer transplants." Gentry and Segev (2015) raise similar concerns motivated by match failures: "...registries forced by competition to perform match runs very

TABLE 1
Fraction of pairs matched and average waiting times in days over all pairs in simulations using NKR data

Arrivals	*Match rate (%)*					*Waiting time in days*				
per day	Greedy	Patient	Batching			Greedy	Patient	Batching		
			7 days	30 days	60 days			7 days	30 days	60 days
0.01	10.7	11.9	10.4	9.9	9.3	322	355	322	324	326
0.05	22.4	23.4	22.2	21.2	20.2	279	324	280	283	288
0.25	34.3	35.4	33.8	32.6	31.2	237	298	238	243	248
0.5	38.5	39.5	38.0	36.8	35.2	222	290	223	228	233
1	42.0	43	41.6	40.2	38.6	209	283	210	215	221
2	45	45.8	44.5	43.1	41.5	198	278	200	205	211
4	47.2	48	46.8	45.4	43.6	190	274	192	196	207

Our theory combined with simulations suggest that matching greedily is not a real source of inefficiency.

1.1. *Related literature*

Closely related literature studies dynamic matching on networks when agents' preferences are based on compatibilities, motivated by kidney exchanges. This literature, initiated by Ünver (2010), can be organized into two perspectives. The first perspective seeks to minimize waiting times in models where agent do not depart exogenously (Ünver, 2010; Anderson, Ashlagi, Gamarnik and Kanoria, 2017; Ashlagi, Burq, Jaillet and Manshadi, 2016). The key finding in this literature is that greedy matching minimizes the average waiting time. The second perspective is concerned with how many agents are matched. Akbarpour *et al.* (2020) consider a model with exogenous departures, in which each agent is compatible with any other agent with a fixed probability. They find that the patient policy leads to an exponentially smaller loss rate (i.e. fraction of unmatched agents) compared to the greedy policy.[6]

Each of these perspectives studies one of two objectives: minimizing the time until an agent is matched, or minimizing the number of agents that leave the market unmatched. The two perspectives lead to different conclusions about the optimality of the greedy policy and suggest a trade-off between matching agents quickly and matching as many agents as possible. This article contributes by studying this trade-off and showing that it vanishes in large kidney-exchange markets with asymmetric agents. Technically, our article is the first to analyse a model with both exogenous departures and heterogeneous agents. We further contribute by analysing the distribution of waiting times rather than just averages and by providing an analysis of finite markets.

The effectiveness of thickening the market by waiting to increase the number of matches has also been studied in markets other than kidney exchanges. Liu, Wan and Yang (2018) compare the match rates of greedy and patient policies in ride-sharing markets for matching drivers with passengers and find that the waiting increases market thickness and average match quality but decreases the number of matches. Finally, recent and indirectly related papers study dynamic matching when agents have preferences beyond compatibility and find that greedy policies are sometimes inefficient since some waiting can improve the quality of matches (Doval, 2014; Ashlagi *et al.*, 2019; Baccara *et al.*, 2020; Mertikopoulos, Nax and Pradelski, 2020; Blanchet, Reiman, Shah and Wein, 2020).

frequently cannot take advantage of mathematical optimization, and likely fewer transplants are accomplished nationwide as a result."

[6] The differences with Akbarpour *et al.* (2020) are discussed in detail in Section 6.

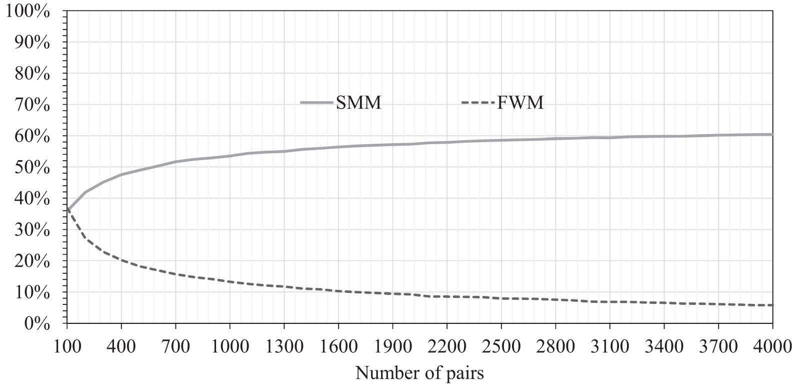


FIGURE 1

Average percentage of pairs without a compatible partner (dashed) and the percentage matched in a maximum matching (solid). The average for every fixed pool size on the horizontal axis is computed by random sampling from the combined data set from NKR, APD, UNOS, and Methodist at San Antonio.

2. THE COMPATIBILITY GRAPH

A kidney exchange pool can be represented by a *compatibility graph* G. Each node in the graph represents an agent (a patient–donor pair), and a link between two nodes exists if and only if the two corresponding agents are *compatible* with each other (so a bilateral exchange between the nodes is feasible). We restrict attention to bilateral exchanges.

A *matching* μ is a set of non-overlapping compatible pairs of agents. Denote by $M(G)$ the set of matchings in G.[7] For every compatibility graph G let $|G|$ denote the number of agents in the graph, and for every matching μ let $|\mu|$ denote the number of agents in that matching.

Define the *(normalized) size of the maximum matching* (SMM) in a graph G to be the fraction of matched agents in a maximum matching:

$$\text{SMM} = \max_{\mu \in M(G)} \frac{|\mu|}{|G|}.$$

Define the *fraction of agents without a partner* (FWP) to be the fraction of agents that are not matched in any matching (thus have no compatible agent):

$$\text{FWP} = \frac{|\{i \in G: \ (i,j) \notin M(G) \text{ for all } j\}|}{|G|}.$$

Figure 1 depicts the expected SMM and FWP for a subset of a given size drawn uniformly at random from the patient–donor population acquired from the National Kidney Registry (NKR), the Alliance for Paired Donation (APD), and the United Network for Organ Sharing (UNOS) and Methodist at San Antonio. These data include 4992 patient–donor pairs.[8] The data allow

[7] The article restricts attention to matching only pairs of agents and not through chains. For the effect of matching through chains see, e.g., Ashlagi, Gilchrist, Roth and Rees (2011) and Anderson *et al.* (2017).

[8] Each data set includes pairs from a different period of time, but no earlier than 2007. The data from NKR, APD and Methodist were obtained directly and is not publicly available. The data from UNOS can be obtained directly from the Organ Procurement and Transplantation Network (OPTN), a contractor for the US Department of Health and Human Services.

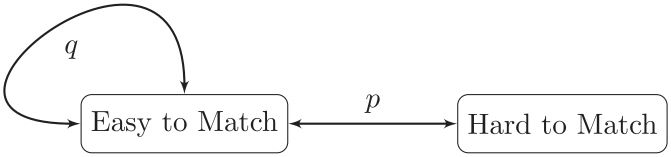


FIGURE 2

The random compatibility model.

to verify whether each patient and donor are virtually compatible, and therefore whether two pairs can match. Two features stand out: first, even as the market grows large, the SMM stays bounded away from 1, i.e., SMM <1. This is a natural consequence of the different blood types (Roth *et al.*, 2007). Second, when the market grows large, the fraction of pairs that have no compatible pairs decreases. Roughly, 5.5% of pairs are incompatible with any other pair in these data (FWP).[9] Since compatibility depends only on the characteristics of the patients and donors, it is independent of pool size, and thus in a sufficiently large pool one would expect that the FWP would further decrease to zero.

Fact 1. *As the kidney exchange patient–donor pool grows large, the compatibility graph (Figure 1) is such that the size of the maximal matching (SMM) stays bounded away from* 1 *and the fraction of patient–donor pairs without a compatible partner (FWP) becomes small.*

The change in both the SMM and FWP measures captures the benefit of a larger market. Since a matching policy in a dynamic environment trades off the benefits of a larger market with the waiting costs incurred by the agents, having a model that accurately represents the SMM and the FWP is important to correctly describe the costs and benefits of waiting to match.

2.1. *A two-type compatibility model*

To capture the features of kidney exchange identified in Fact 1, we adopt a stylized and tractable model with random compatibilities. There are two types of agents, *easy-to-match* or *hard-to-match*, denoted by E and H, respectively. There are more hard-to-match than easy-to-match agents. Any pair of hard-to-match and easy-to-match agents are compatible independently with probability $p>0$, any pair of easy-to-match agents are compatible independently with probability $q>0$, and no pair of hard-to-match agents are compatible with each other (Figure 2).

Proposition 1 shows that this simple model is indeed able to capture the features of real kidney exchanges identified in Fact 1.

Proposition 1. *Consider a compatibility graph with m easy-to-match agents and* $(1+\lambda)m$ *hard-to-match agents where* $\lambda>0$. *Compatibilities between pairs of agents are generated as described in Section 2.1. Then, as m grows large* $\text{SMM}=\frac{2}{2+\lambda}$ *and* $\text{FWP}=0$ *hold with high probability.*[10]

That the size of the maximal matching cannot exceed $\frac{2}{2+\lambda}$ is intuitive: since H agents cannot match with each other and there are more H agents than E agents, some H agents must remain

[9] In practice, some patients can receive a kidney from blood-type incompatible donors due to advanced technology. For the sake of simplicity, we ignore this in our simulations, but it is worth noting that the FWP drops to less than 3% when this form of compatibility is allowed.

[10] We say a sequence of events $E_1,E_2,\ldots$ hold with high probability when $\lim_{k\to\infty}\mathbb{P}[E_k]=1$.

unmatched when the pool is large. An upper bound on the fraction of agents that can be matched equals twice the fraction of E agents $\frac{1}{2+\lambda}$. Furthermore, note that this fraction is achieved whenever there exists a matching in which all E agents are matched with H agents. It follows from a standard result in random graph theory that the probability that such a "perfect matching" exists approaches 1 as the pool grows large. Furthermore, as the pool grows large, any H agent will be compatible with some E agent, since compatibilities between agents are drawn independently. Thus, the FWP converges to 0. The parameter λ of the model measures the degree of imbalance between hard- and easy-to-match agents. So $\lambda=0$ corresponds to a balanced pool. Figure 1 suggests that the size of the maximal matching in the data converges to roughly 60% when the pool becomes large, implying that $\lambda \approx 1.33$ in the context of our model.

Proposition 1 establishes that our two-type model can match the empirical behaviour of the SMM and FWP measures observed in Fact 1. Proposition 2 establishes that no model with a single type can replicate the empirical features of real kidney exchanges that the size of the maximal matching is less than one while the FWP goes to zero, even when allowing the probability of compatibility between two agents to depend on the market size in arbitrary ways.

Proposition 2. *Consider a model with m homogeneous agents, in which every pair of agents are compatible independently with probability* $p(m)>0$ *that may depend on the market size. The following two conditions cannot be satisfied simultaneously:*

$$\lim_{m\to\infty} \mathbb{E}[\text{SMM}] < 1, \textit{and} \tag{1}$$

$$\lim_{m\to\infty} \mathbb{E}[\text{FWP}] = 0. \tag{2}$$

The proof is constructive. It begins with assuming that every agent has a compatible partner when the pool grows large, i.e., (2) is satisfied. It then introduces an algorithm which constructs a matching that includes all agents with high probability as the pool grows large. This implies that (1) and (2) cannot hold together in *any* random graph model with homogeneous agents.

Economically, this observation implies that heterogeneity of agents plays a major role in kidney exchanges.[11] Our two-type model is arguably the simplest random compatibility model that captures these features of the compatibility graph. It may be useful to illustrate the connection with kidney exchanges while restricting attention to blood-type compatibilities. One may think of A–O and O–A patient–donor pairs as easy- and hard-to-match, respectively. More generally patient–donor pairs with blood types X–O for X≠O, AB–A, and AB–B are blood-type compatible with a pair of the same category and can be considered as easy-to-match, whereas those with blood types O–X for X≠O, A–AB, and B–AB are not, and hence can be considered as hard-to-match. Typical exchange pools have fewer easy-to-match than hard-to-match pairs simply because a patient who is compatible with her intended donor will be transplanted directly. For example, an A patient and her intended O donor may be compatible and would not join the exchange (they only need to join if they are tissue-type incompatible).

3. DYNAMIC MATCHING

We embed the static compatibility model from Section 2.1 in a dynamic model that allows to study matching policies in a dynamic setting. We consider an infinite-horizon dynamic model, in which

[11] This is consistent with Roth *et al.* (2007) and Agarwal, Ashlagi, Azevedo, Featherstone and Karaduman (2019), who demonstrate that the types of patients and donors play a crucial role for efficiency.

agents can match bilaterally. Easy-to-match agents arrive to the market according to a Poisson process with rate m, and hard-to-match agents arrive to the market according to an independent Poisson process with rate $(1+\lambda)m$. We assume that the majority of agents are hard-to-match, that is $\lambda > 0$, unless explicitly stated otherwise. We sometimes refer to m as the *market size*.

An agent that arrives to the market at time t becomes *critical* after Z units of time in the market, where Z is distributed exponentially with mean d, independently between agents. We refer to $1/d$ as the *criticality rate*. The latest an agent can match is the time she becomes critical, $t+Z$; immediately after this time the agent leaves the market unmatched. The planner observes when an agent gets critical and can attempt to match the agent immediately at that time before she departs.[12]

Matching policies. Denote by G_t the compatibility graph induced by the agents that are present at time t. A *dynamic matching policy* selects at any time t a matching $\mu_t \in M(G_t)$, which may be empty. Whenever a non-empty matching is selected, all matched agents leave the market.

Several kidney exchange platforms in the US attempt to match pairs as soon as they arrive to the market (Ashlagi *et al.*, 2018). A tractable approximation of this behaviour is a greedy policy.

Definition 1 (Greedy) *In the* greedy *policy, an agent is matched upon arrival with a compatible agent if such an agent exists. If she is compatible with more than one agent, H agents are prioritized over E agents and otherwise ties are broken randomly.*

Some platforms identify matches only periodically, allowing the pool to thicken and possibly offer more matching opportunities. For example, UNOS matches twice a week, whereas national platforms in the UK and the Netherlands identify matches every 3 months (Biro, Burnapp, Bernadette, Hemke, Johnson, van de Klundert and Manlove, 2017). This behaviour is approximated with the following batching policy.

Definition 2 (Batching) *A* batching *policy executes a maximal match every T units of time. If there are multiple maximal matches, select randomly one that maximizes the number of matched H agents. The parameter T is called the* batch length.[13]

We also consider the patient policy, proposed by Akbarpour *et al.* (2020), which attempts to match an agent only once she becomes critical. In the context of kidney exchange, this means that two patient–donor pairs in the pool are matched only if the condition of one of these pairs is such that it cannot match at a later point in time.

Definition 3 (Patient) *In the* patient *policy, an agent that becomes critical is matched with a compatible agent if one exists. If she is compatible with more than one agent, H agents are prioritized over E agents, and otherwise ties are broken randomly.*

The patient policy can be viewed as a theoretical benchmark, as predicting the time at which the patient will become too sick to transplant is generally not feasible. Observe that the greedy

[12] As we will show that the asymptotically optimal policy does not condition on this information, thus allowing for this larger set of policies strengthens our results.

[13] We note that every agent leaves the market either because of being matched to another agent or because of getting critical. Thus, in the batching policy, (only) the agents who are in an executed matching are removed from the market at the time the matching is executed.

and patient policies match at most two agents at any given time because no two agents ever arrive or become critical at the same time. The batching policy, however, can match multiple agents when it executes a matching.

Measures for performance. Denote by $\theta_i \in \{E,H\}$ agent i's type, by $\alpha_i \geq 0$ her arrival time, by $\varphi_i \geq 0$ how long she is present in the market, and indicate by $\mu_i \in \{0,1\}$ whether she is matched. To study the performance of a matching policy, we focus on two measures. One is the *match rate* $q_\Theta(m) \in [0,1]$ of each type $\Theta \in \{E,H\}$, which is the fraction of agents of type Θ that get matched. Formally, we define the match rates for each arrival rate m by

$$q_\Theta(m) = \lim_{t\to\infty} \mathbb{E}\left[\frac{|\{i\colon\, \mu_i = 1 \text{ and } \alpha_i \leq t \text{ and } \theta_i = \Theta\}|}{|\{i\colon\, \alpha_i \leq t \text{ and } \theta_i = \Theta\}|}\right].$$

The other is the *expected waiting time* (or simply *waiting time*) $w_\Theta(m)$ of agents of type Θ, whether eventually matched or not. Formally, we define the waiting time for each arrival rate m by

$$w_\Theta(m) = \lim_{t\to\infty} \mathbb{E}\left[\frac{\sum_{i\colon\, \alpha_i \leq t \text{ and } \theta_i = \Theta} \varphi_i}{|\{i\colon\, \alpha_i \leq t \text{ and } \theta_i = \Theta\}|}\right].$$

One reason for studying match rate and waiting time is that they together determine the payoff of a risk-neutral expected-utility-maximizer (EU) who assigns a fixed value to being matched and incurs a constant cost while waiting in the market.

We are interested in optimal policies for large pools and denote by $q_\Theta = \lim_{m\to\infty} q_\Theta(m)$ and $w_\Theta = \lim_{m\to\infty} w_\Theta(m)$ the match rate and waiting time when the market becomes (infinitely) large, i.e., when the arrival rate m goes to infinity.[14] We consider the following notion of optimality:

Definition 4 (Asymptotic optimality) *A policy is asymptotically optimal if for every $\epsilon > 0$ there exists m_ϵ such that, when $m \geq m_\epsilon$,* no *type of agent can improve its match rate $q_\Theta(m)$ or expected waiting time $w_\Theta(m)$ by more than ϵ when changing to any other policy.*

This optimality notion is demanding, since it requires the policy to be optimal for every type of agent simultaneously. It is unclear whether an asymptotically optimal policy exists, since a policy that is optimal for H agents might be suboptimal for E agents.

3.1. *Results*

In this section, we present a characterization of the match rates and waiting times associated with the greedy, batching and patient matching policies and discuss its implications.

Theorem 1. *The greedy policy is asymptotically optimal, whereas the batching policy (for any fixed batch length) and the patient policy are not asymptotically optimal.*

We further compute the match rates and expected waiting times under these policies.

Proposition 3. *As the arrival rate m grows large:*

[14] Throughout, we restrict attention to policies where these limits are well defined. This assumption could be relaxed by considering the liminf and limsup in the definitions.

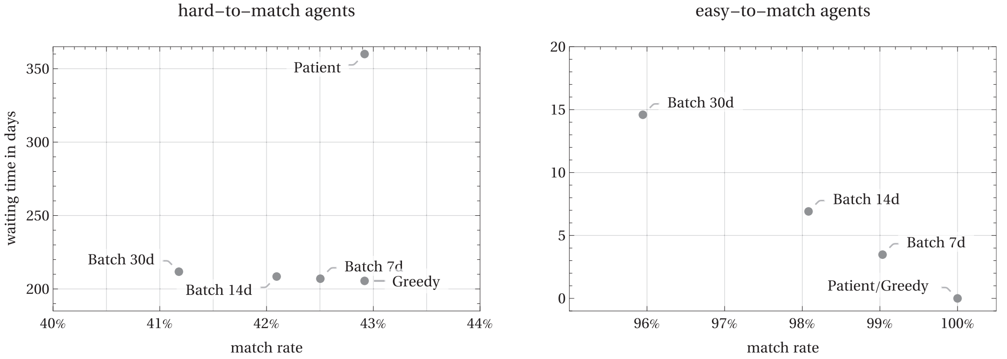


FIGURE 3

Illustration of Proposition 3 when $\lambda = 1.33$ and d equals 360 days. The points represent the predictions of our model for large markets, which are derived in Proposition 3.

(i) The match rates of hard- and easy-to-match pairs under the greedy policy are $\left(q_H^G, q_E^G\right) = \left(\frac{1}{1+\lambda}, 1\right)$*, respectively, and their expected waiting times are* $\left(w_H^G, w_E^G\right) = \left(\frac{\lambda d}{1+\lambda}, 0\right)$.

(ii)A batching policy with batch length T achieves match rates of $\left(q_H^B, q_E^B\right) = \left(\frac{1-e^{-T/d}}{(1+\lambda)T/d}, \frac{1-e^{-T/d}}{T/d}\right)$. *Furthermore, the expected waiting time for each type* Θ *is* $w_\Theta^B = d(1-q_\Theta)$. *Also,* $q_\Theta^B < q_\Theta^G$*, whereas* q_Θ^B *approaches* q_Θ^G *as T approaches* 0. *In addition,* $w_\Theta^B > w_\Theta^G$*, whereas* w_Θ^B *approaches* w_Θ^G *as T approaches* 0.

(iii)The match rates of hard- and easy-to-match pairs under the patient policy approach $\left(q_H^P, q_E^P\right) = \left(\frac{1}{1+\lambda}, 1\right)$*, respectively, and their expected waiting times approach d and* 0*, respectively.*

Figure 3 illustrates the match rates and waiting times of H and E agents under the different policies as found in Proposition 3. In the figure, the values for λ, d are chosen to match the imbalance and criticality rate in the NKR data ($\lambda = 1.33, d = 360$). As Figure 3 illustrates, each batching policy leads agents to wait longer and get matched with a smaller probability than under the greedy policy. For example, in comparison to greedy, under a monthly batching policy hard-to-match agents wait on average 6 days longer and easy-to-match agents 15 days longer. Hard- and easy-to-match agents get matched with 1.7% and 4% lower probability. Similarly, the patient policy matches equally many agents as the greedy policy but leads to a substantially longer expected waiting time for hard-to-match agents (155 more days).

Remark 1. *It is important to note that Theorem 1 and Proposition 3 do not imply that batching policies are suboptimal for a fixed market size. For a fixed market size, a batching policy which matches very frequently will achieve (almost) the same outcome as the greedy policy and thus will also be close to optimal in large markets.*[15] *We investigate this in detail in Section 4 and show how frequent batching policies must match to be close to optimal.*

We now provide rough intuition for the differences between greedy, batching and patient matching policies. In Section 3.2, we sketch the argument for the various parts of the results. Section 4 provides a more extensive comparison of these three policies, including a non-asymptotic analysis.

[15] We formally establish this in Proposition 6.

Intuition for the optimality of greedy. As there are more hard- than easy-to-match agents, hard-to-match agents will accumulate and a large number of them will be waiting to be matched at any time under any policy.[16] This implies that under greedy matching, easy-to-match agents will have upon arrival, with high probability, a compatible hard-to-match agent and are therefore matched immediately. So every easy-to-match agent is matched with a hard-to-match agent, implying that greedy matching asymptotically achieves the optimal match rate. Intuitively, the market for hard-to-match agents already thickens under the greedy policy and further thickening the market is not beneficial for increasing the match rate.

Under the batching policy, each agent waits at least from the time of her arrival until the next time a matching is identified. Thus each agent waits on average at least half the length of the batching interval. Furthermore, each agent becomes critical during that time with strictly positive probability. Thus, easy-to-match agents are worse off under the batching policy than under the greedy policy where they get matched immediately with probability 1. As some easy-to-match agents leave the market unmatched, hard-to-match agents are matched with a smaller probability. Consequently there are, on average, more hard-to-match agents waiting in the market. Little's law, which states that the arrival rate multiplied by the average waiting time equals the average number of waiting agents (Little and Graves, 2008), implies that hard-to-match agents also wait longer under any batching policy than under a greedy matching policy. As both types are worse off, batching policies are not asymptotically optimal.

Under the patient policy, so many hard-to-match agents accumulate that an easy-to-match agent will, with high probability, match with a critical hard-to-match agent almost immediately upon arrival. This implies that the policy asymptotically achieves the optimal match rate. As hard-to-match agents get matched only when they become critical, the distribution and expectation of their waiting time is the same as if they do not match at all. Hence, hard-to-match agents get matched faster under a greedy policy, implying that the patient policy is not asymptotically optimal.

Intuitively, when the departure rate is small, increasing the market size and making the market thicker by waiting are loosely speaking substitutes in the sense that they increase the match rate. This is an intuition for why the greedy policy becomes optimal when the market grows large and the benefit of thickening the market vanishes. We note that thickening the market through waiting and increasing the market size exogenously lead to quite different compositions of waiting agents. If the arrival rate increases, under a greedy policy, only hard-to-match agents will accumulate (as in a large market easy-to-match agents will match immediately). In contrast, if the market is thickened by waiting in a batching policy, both hard- and easy-to-match agents accumulate.

3.2. *Discussion of results*

In this section, we provide a proof sketch for the various parts of Theorem 1 and Proposition 3 as well as additional results on the waiting time distributions. We first establish an upper bound on the performance of *any* policy.

Proposition 4 (Upper bound on the performance of any policy) *For any market size m, and under any policy, the match rate of hard-to-match agents is at most $\frac{1}{1+\lambda}$ and their expected waiting time is at least $\frac{\lambda d}{1+\lambda}$.*

[16] On the other hand, if there are more easy- than hard-to-match agents, then under the greedy policy only a small number of hard-to-match agents would be in the pool at any time; because otherwise, arriving easy-to-match agents will often be matched to hard-to-match agents, which will reduce the number of hard-to-match agents in the pool over time.

The result on the match rate follows from the fact that H agents cannot match with other H agents. The arrival rate of E agents is only m per unit of time, compared to $(1+\lambda)m$ for H agents. Thus, the strong law of large numbers implies that a fraction of $\frac{\lambda}{1+\lambda}$ of the H agents remain unmatched in the long run. The result on waiting times is less immediate. We prove it by first deriving a policy-specific lower bound on the match rate; this lower bound holds with equality in policies where every agent who gets critical leaves the market unmatched.[17] Combining this lower bound on the match rate with the previously derived upper bound gives us a lower bound on the average number of H agents present in the market, which, through an application of Little's law, yields a lower bound on their waiting time.

3.2.1. Greedy policy.

Next, we analyse the performance of the greedy policy as the market grows large. The following proposition includes the results in the first part of Proposition 3.

Proposition 5 (Performance of the greedy policy) *Consider the greedy policy as the market grows large, i.e.,* $m \to \infty$*. The match rate of* H *agents converges to* $\frac{1}{1+\lambda}$ *and their waiting time converges to an exponential distribution with mean* $\frac{\lambda d}{1+\lambda}$*. The match rate of* E *agents converges to* 1 *and their waiting time converges to* 0.

We first provide intuition for the waiting time distribution. Consider greedy matching in a deterministic setting where every E agent is compatible with every H agent, agents arrive deterministically, and get critical after precisely d units of time. In this setting, E agents will be matched upon arrival with H agents. This means that there will be no E agents waiting in the market, and their waiting time equals zero. Denote by x the steady-state number of H agents present in the market. Per unit of time, the number of H agents arriving to the market equals $(1+\lambda)m$, and m of them are matched with E agents. Furthermore, $\frac{x}{d}$ of the waiting agents are expected to depart unmatched per unit of time. At the steady state, the number of unmatched departing agents equals the number of unmatched arriving agents. Thus, x solves the balance equation

$$\frac{x}{d} = \lambda m \Rightarrow x = \lambda m d. \tag{3}$$

Therefore, if the matching partner for an E agent is chosen at random, each H agent is matched at rate $\frac{m}{\lambda m d} = \frac{1}{\lambda d}$. The time at which a never-departing H agent would be matched is therefore exponentially distributed with rate $\frac{1}{\lambda d}$. The time until an H agent becomes critical is exponentially distributed with rate $1/d$. Since, the minimum of two exponentially distributed random variables is again exponentially distributed with rate equal to the sum of the rates, the waiting time of an H agent is exponentially distributed with the rate $\frac{1+\lambda}{\lambda d}$, and thus with mean $\frac{\lambda d}{1+\lambda}$.

The formal proof of Proposition 5 is more complex as it needs to deal with the randomness in compatibilities, arrivals, and criticality times. Our analysis is based on the Lyapunov function method.[18] In our case, a Markov chain tracks the number of easy- and hard-to-match agents in the

[17] Thus, the lower bound holds with equality in the greedy and batching policies, but not in the patient policy.

[18] Variations of this method are widely used to identify stable points of ordinary differential equations and to analyse steady states of stochastic systems (see e.g. Khalil, 2009; Brémaud, 2013). The idea is defining a function on the state space of a Markov chain such that the expected change of the function is negative outside of a "small box" and possibly positive inside that box. The fact that the expected change of the function equals zero at the steady-state distribution implies a bound on the time the process can spend outside the box. This idea can be used to provide bounds on the expectation of a given function f defined over a Markov chain (Anderson *et al.*, 2017).

pool, and the Lyapunov function argument translates into a concentration bound for the number of easy- and hard-to-match agents in the pool. The number of H agents in the pool at any time is, with high probability, not more than an additive factor of $\sqrt{m}\log m$ away from the solution of the balance equation (3). As this distance grows slow relative to the market size, the Markov chain is well-approximated by the dynamics of the deterministic setting described earlier.

3.2.2. Batching policy. We next sketch the analysis for the match rates and waiting times under the batching policy, as given in the second part of Proposition 3. We start by providing lower and upper bounds on the match rate of E agents. A simple upper bound on the match rate can be derived based on the fact that an arriving agent should wait until the next matching period and may not get matched if she becomes critical before that. We compute $\gamma_{T,d} = \frac{1-e^{-T/d}}{T/d}$ as the probability that an agent does not become critical before the first matching period after her arrival. This is clearly an upper bound on the match rate of E agents. Then, from the fact that H agents are compatible only with E agents, we imply that the match rate of H agents is at most $\frac{\gamma_{T,d}}{1+\lambda}$.

Providing a lower bound on the match rate of E agents is more involved. The key idea is showing that, every time when a matching is executed, the number of H agents matched is at least the number of E agents who are present in the pool at that time and arrived after the previous matching execution. The proof for this fact is based on a probabilistic analysis argument and uses *augmenting path* techniques from Berge's lemma in matching theory[19]. By this fact, almost every E agent who participates in at least one execution of the matching is matched to an H agent. This translates into an asymptotic lower bound of $\gamma_{T,d}$ on the match rate of E agents, and an asymptotic lower bound of $\frac{\gamma_{T,d}}{1+\lambda}$ on the match rate of H agents. Finally, the claim about the match rates follows immediately from the fact that the upper and the lower bound coincide.

To compute expected waiting times, we first note that $y/d = m(1-q_E(m))$, where y is the time-average number of E agents in the pool and $1-q_E(m)$ is the probability of an E agent not getting matched. This holds because, under the batching policy, the number of E agents that get critical equals the number of E agents that do not get matched. We then note that the probability $1-q_E(m)$ converges to $1-\gamma_{T,d}$ as m approaches infinity, since the match rate of E agents is $\gamma_{T,d}$. Thus, $d(1-\gamma_{T,d}) = \frac{y}{m}$. The right-hand side is the expected waiting time of E agents, by Little's law. Hence, the expected waiting time of E agents equals $d(1-\gamma_{T,d})$. A similar argument proves the claim for the waiting time of H agents.

Batching with vanishingly batch length. We next strengthen our previous analysis and show that a batching policy is asymptotically optimal if and only if the batch length vanishes with the market size.

Proposition 6. *A market size dependent batching policy with batch length T_m is asymptotically optimal if and only if the batch length goes to zero as the market becomes large* $\lim_{m\to\infty} T_m = 0$.

The intuition for the only if direction is that if $T_m > \delta > 0$ for every m, then the probability that a newly arrived agent becomes critical before a matching is executed is positive and does not vanish as the market grows large. Thus, the match rate of E agents would be below their match rate under the greedy policy. To prove the if direction, one cannot just take the limit $T \to 0$ of the expressions for match rate and waiting time obtained in Proposition 3, since the limits with respect to time and market size are not interchangeable. In particular, as T_m approaches 0, the

[19] See, e.g., West (2000) for more on Berge's lemma and augmenting paths.

Poisson concentration bounds on the number of arrivals in each period that we use to establish Proposition 3 become arbitrarily weak. Instead, we show that there is always a large number of H agents waiting in the pool, and an arriving E agent would be matched to one of these agents with high probability.

3.2.3. Patient policy.

We next quantify the performance of the patient policy. The following proposition includes the third part of Proposition 3.

Proposition 7 (Performance of the patient policy) *Consider the patient policy when the pool grows large, i.e., $m \to \infty$. The match rate of H agents converges to $\frac{1}{1+\lambda}$ and their waiting time converges to an exponential distribution with mean d. The match rate of E agents converges to* 1 *and their waiting time converges to* 0.

To get some intuition for this result, consider again the hypothetical case in which every H agent is compatible to every E agent, and agents arrive and get critical deterministically. In steady state, there are almost no E agents in the market and the number of H agents in the market is approximately $(1+\lambda)md$. To see why, suppose this is indeed the state of the market at the beginning of time. H agents get critical and attempt to match with E agents at a rate of $\frac{(1+\lambda)md}{d} = (1+\lambda)m$. Since this rate is much larger than the arrival rate of E agents m, then E agents are matched almost immediately at the steady state. Thus, their number remains close to zero at any time. Consequently, almost no E agent becomes critical, and almost all matches are initiated due to an H agent becoming critical. Since H agents arrive at rate $(1+\lambda)m$ and get critical at rate $1/d$, their number in the pool remains close to $(1+\lambda)md$, and the steady state is maintained. As H agents are the ones that initiate matches, their average waiting time equals the average time until they become critical, d.

4. COMPARISON OF BATCHING AND GREEDY POLICIES IN FINITE MARKETS

Our results imply that, in a large market, greedy will outperform batching if the length of the batching interval does not go to zero (Theorem 1 and Proposition 6). It is natural to ask how frequent batching policies must execute matches to (potentially) outperform the greedy policy in a *finite* market. We address this question in two ways. First, we derive an upper bound on the batch length such that *any* batching policy with a larger batch length matches fewer agents and has a higher waiting time than the greedy policy. Second, we run simulations based on the model and real data that indicate that at realistic market sizes and parameters in the context of kidney exchanges, greedy matching dominates batching policies that match less frequently than daily.

4.1. *An analytical bound on the performance of batching in finite markets*

The next result combines non-asymptotic upper bounds on the performance of batching policies with lower bounds on the performance of the greedy policy to establish a bound on the batch length such that less frequent batching will be dominated by the greedy policy in a *finite* market.

Proposition 8. *Let $m > 0$ be an arbitrary* fixed *arrival rate. Define z^* to be the steady-state probability that an easy-to-match agent, upon her arrival, is matched to a hard-to-match agent under the greedy policy. Then, for every agent type (easy- and hard-to-match), the match rate and waiting time of that type under the batching policy are respectively smaller and larger than*

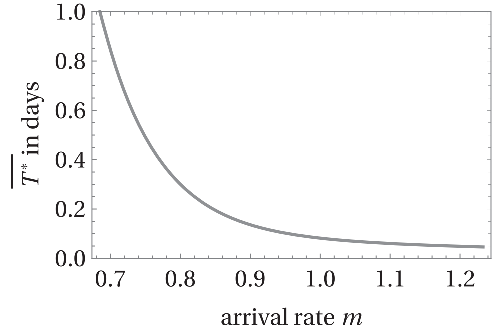


FIGURE 4

The batch length above which greedy dominates batching for various arrival rates per day, $\lambda = 1.33, p = 0.037$, and average criticality time $d = 360$ days. The bound $\overline{T^*}$ is independent of $q \in [0,1]$ and is decreasing in p.

under the greedy policy if the batch length T satisfies $T > T^*$, where

$$T^* = \frac{z^* W\left(-\frac{e^{-1/z^*}}{z^*}\right)+1}{z^*/d} \tag{4}$$

and $W(\cdot)$ is the Lambert W function.[20]

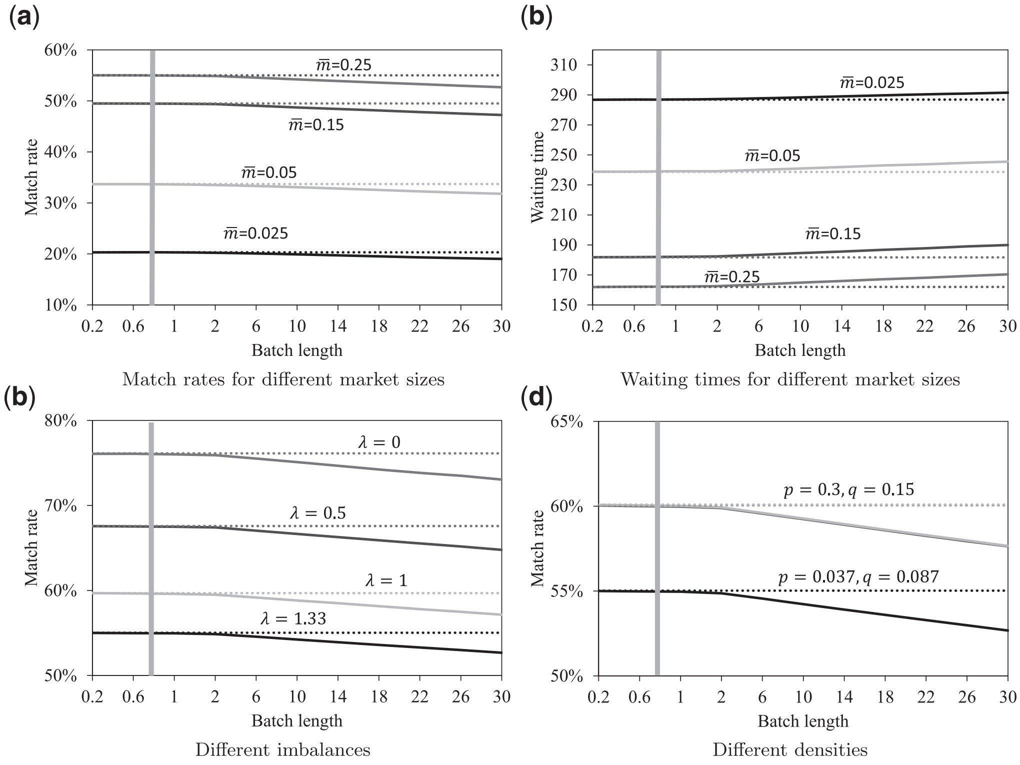
The above proposition involves the exact value for the steady-state probability of an easy-to-match agent being matched to a hard-to-match agent upon her arrival. In Supplementary Appendix ii, we compute a lower bound on this probability using the Lyapunov function method and combine it with the above result to get an upper bound for T^*, namely $\overline{T^*}$. Therefore, a batching policy that makes matches less frequently than $\overline{T^*}$ is dominated by the greedy policy for the *fixed* arrival rate m. Figure 4 visualizes this bound. Recall that $\lambda = 1.33$ is consistent with the NKR data (see Section 2.1). Using the same data, we can calibrate the compatibility probabilities between the types by defining easy-to-match agents as those who are part of every maximal matching, which leads to empirical compatibilities of $p \approx 0.037, q \approx 0.087$.[21] For these parameters, matching less frequently than daily leads to a lower match rate than greedy for any arrival rate of $m \geq 0.7$ per day (roughly twice the size of the NKR). We note that Proposition 8 provides a sufficient but not necessary condition. The next section uses simulations to explore smaller arrival rates.

4.2. *Numerical simulations*

We run simulations based on our two-type model to compare greedy and batching for various batch lengths and different parameters. The first set of simulations varies the total arrival rate per day $\bar{m} = m + (1+\lambda)m$, the imbalance λ and compatibility structure (p,q). In line with our calibration to the NKR data, the base case values are set to $\lambda = 1.33$ and $(p,q) = (0.037, 0.087)$.

[20] We recall that the Lambert W function is the inverse of the function $F(w) = we^w$.

[21] It is important to note that the patient–donor pairs identified as hard-to-match are not necessarily blood-type incompatible. Typically, patients of blood-type compatible pairs are much more sensitized, which means they have a lower chance to match with a random blood-type compatible donor. As there are few donors who can match highly sensitized patients, highly sensitized blood-type compatible pairs are typically not part of every maximal matching and our calibration thus identifies them as hard-to-match. The simulation results that use these calibrations are also confirmed by simulations that directly use real compatibility data.



Match rates for different market sizes

Waiting times for different market sizes

Different imbalances

Different densities

FIGURE 5

Comparison of the greedy (dashed line) and the batching (solid line) policies in terms of the fraction of matched agents and waiting times. Batch lengths are in days. Comparisons are plotted for different total arrival rates $\bar{m}$ (a, b), different imbalance values λ (c), and different compatibilities (d).

The *total arrival rate* (sum of arrival rates of both types) of agents is set to $\bar{m} = 0.25$, i.e., an agent every 4 days (and thus lower than the arrival rate at the NKR). Agents become critical on average after 360 days. This exercises suggests that even in moderately sized markets batching needs to be very frequent for the batching policy not to be outperformed by the greedy policy. The results are reported in Figure 5. In each of these simulations, a batching policy which matches less frequently than once a day will result in a lower match rate than greedy while matching more frequently than daily will be indistinguishable from greedy matching.[22]

Similar results hold for a wide range of parameters in the two-type model that are consistent with kidney exchange. A natural question is whether it is the case for *all* parameterizations of the model that greedy matching dominates batching in the two-type model. This turns out not to be true and we next provide a counter example.

An example where batching is beneficial. We identify a non-asymptotic setting where batching leads to a higher match rate than greedy. Consider the scenario in which all easy-to-match agents are compatible $q = 1$ and hard- and easy-to-match agents are rarely compatible

[22] Figure 5d also plots results for values of $(p,q) = (0.3, 0.15)$ that are chosen to match the empirical frequency with which blood-type compatible and incompatible pairs can match.

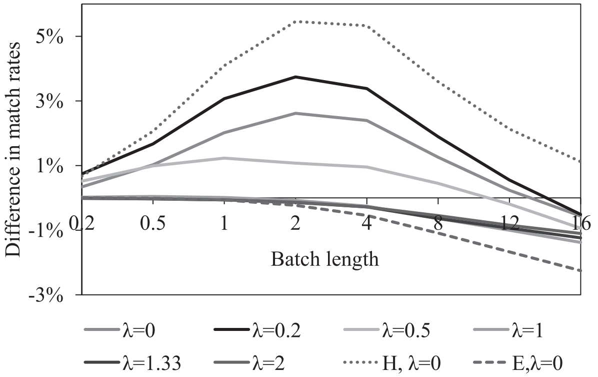


Figure 6

The difference between the match rates of batching and greedy policies for extreme match probabilities $p=0.02, q=1$, and various market imbalances.

$p=0.02$.[23] The total arrival rate equals $\bar{m}=2$ agents per day and agents become critical after $d=360$ days. Intuitively, as in this market many easy-to-match agents have no hard-to-match partner, the greedy policy frequently matches easy-to-match agents to themselves; this results in more unmatched hard-to-match agents compared to the batching policy which matches fewer easy-to-match agents to themselves. As shown in Figure 6, when the market is balanced ($\lambda=0$) a batching policy with a batch length of two days leads to matching about 2.6% more agents than greedy. The parameters in this example are carefully chosen and the benefit from batching vanishes when either the market size grows large—as predicted by Theorem 1—or as the market imbalance grows large.

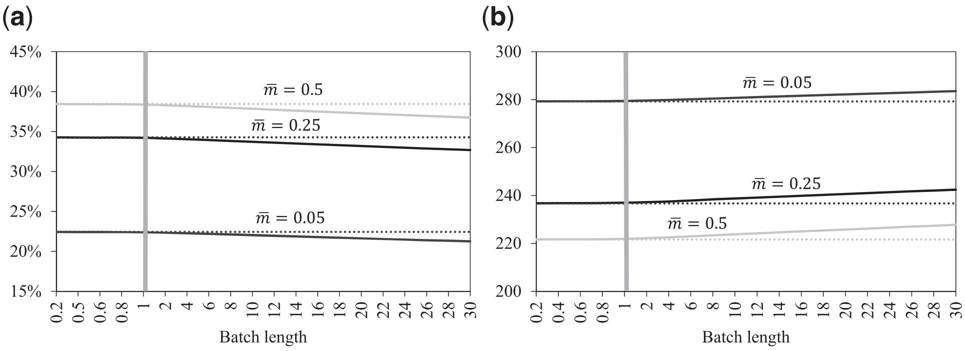
The effect of imbalance. To see why imbalance has an effect on the optimality of the greedy policy, note that the number of H agents in the pool under greedy is less than $\lambda md - k\sqrt{md}$ with probability at most $O(e^{-k})$.[24] When the number of H agents is at least $\lambda md - k\sqrt{md}$, the probability that an arriving E agent is not matched to an H agent is at most $(1-p)^{\lambda md - k\sqrt{md}}$. A union bound then implies that the probability that an arriving E agent is not matched to an H agent is at most of the order of $e^{-k}+e^{-p(\lambda md - k\sqrt{md})}$ for every $k>0$. Setting $k=\lambda\sqrt{md}/2$ implies that this probability is of the order of $e^{-\lambda\sqrt{md}/2}$, which approaches zero exponentially fast in $\lambda\sqrt{md}$. So, the inefficiency that results from E agents being matched to each other in the greedy policy vanishes if the market is either large (m is large) or imbalanced (λ is large).

5. EMPIRICAL FINDINGS

While the greedy policy is optimal in a sufficiently large or imbalanced market, it is ultimately an empirical question whether our results hold in practice. We next complement our theoretical predictions with simulations using kidney exchange data. These simulations indicate that the

[23] These parameters are not in line with kidney exchange data as blood-type compatible patient–donor pairs are selected to have highly sensitized patients and therefore typically have a low chance to be compatible to each other.

[24] This follows from our concentration bounds; see Theorem 4 in the Appendix.



Match rate

Waiting time

FIGURE 7

Match rate (left) and average waiting times (right) under greedy (dashed line) and batching (solid line) policies. Batch length given in days. $\bar{m}$ is the total arrival rate per day.

greedy policy dominates batching in terms of match rate and waiting time and have a much better waiting time compared to patient and a slightly worse match rate.[25]

The simulations presented here use data from the NKR. The NKR data include 1881 de-identified patient–donor pairs between July 2007 to December 2014.[26] We use patients' and donors' blood types, antigens and antibodies to verify (virtual) compatibility between each donor and each patient. On average, approximately one patient–donor pair arrives per day to the NKR, and the average criticality time of a pair is estimated to be 360 days.[27] Arrivals of pairs are generated according to a Poisson process with a fixed arrival rate. We vary the arrival rate capturing market sizes from one-tenth to four times the size of the NKR. Pairs become critical according to an independent exponentially distributed random variable with mean equal to 360 (days), based on the empirical estimate. We simulate greedy, patient, and batching policies until 10 million pairs arrive to the market and report match rate and waiting time by taking averages over all or over a predefined subset of pairs of a certain type.

Table 1 reports the fraction of matched pairs and average waiting time. For the batching policy, we report results for weekly, monthly, and bimonthly batching ($T=7,30,60$ days). The patient policy always results in the highest match rate, and the greedy and weekly batching result in a slightly lower match rate (and larger batch lengths result in a lower match rate). Moreover, the average waiting time under greedy matching is the smallest among all policies.

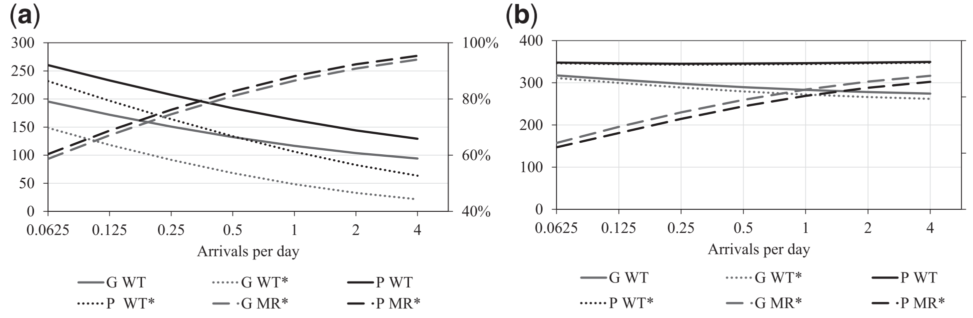
Next, we compare the greedy policy and batching policies with a finer range of batch lengths. The results are plotted in Figure 7 for three different total arrival rates ($\bar{m}$). In all cases, the greedy policy outperformed the batching policy in match rate and waiting time.

In the next simulations, we address the concern that some types might be harmed by the greedy policy. To do so, we compute average waiting times and match rates separately for two types of pairs, *under-demanded* and *over-demanded*. These can be thought of as hard- and easy-to-match, respectively. Under-demanded patient–donor pairs are blood type incompatible with each

[25] Note here that the patient policy constitutes a theoretical benchmark as it is often impossible to observe criticality. The greedy and batching policies do not use any criticality information.

[26] Our focus is on bilateral matching and we therefore omit altruistic donors from the data.

[27] Hazard rates vary only slightly across pair types, such that for the sake of simplicity we aggregate all pairs and use a simple hazard rate model from Agarwal *et al.* (2019) to estimate criticality rate.



Over-demanded pairs

Under-demanded pairs

FIGURE 8

Average waiting times (WT) and match rate (MR) in days under greedy (G) and patient (P) policies. The left and right axes are WT and MR. The label (*) excludes pairs who have no match in the data.

other (these include blood types patient–donor pairs O–X for X≠O, A–AB, and B–AB).[28] Over-demanded pairs are blood type compatible with each other (but not tissue-type compatible) and include pairs X–O for X≠O, AB–A, and AB–B.[29] Figure 8 reports the results. The waiting times (solid lines) of over-demanded pairs steadily decrease as the market becomes thicker, whereas the average waiting times of under-demanded pairs change only slightly. This finding is in line with the predictions from Proposition 3. Despite the heterogeneity in the data, the theoretical predictions (of the stylized two-type model) are aligned with the experiments when we categorize pairs as either over-demanded or under-demanded. These patterns hold even though patients belonging to over-demanded pairs are, on average, more sensitized than those in under-demanded pairs.[30] We also report the statistics for the set of pairs that have at least one match in the historical data (dotted lines labelled with *).

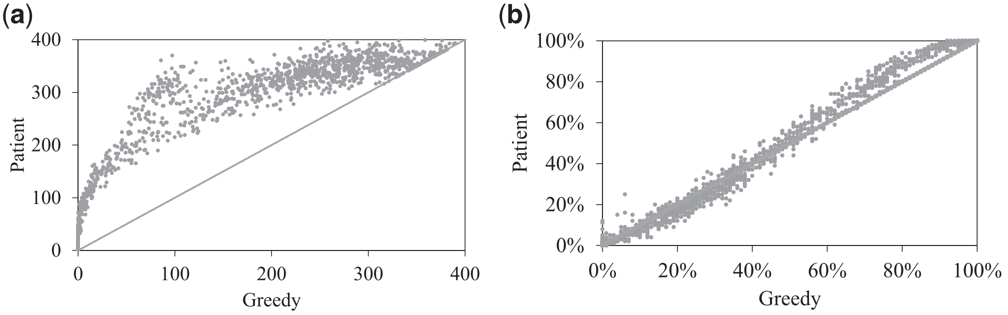
In the last simulation, we run greedy and patient policies under the base case scenario (with an arrival rate of 1 pair per day). For each pair, we compute the average waiting time over the copies of this pair sampled in the simulation as well as the fraction of the copies that are matched (i.e. the empirical probability of getting matched). For each of the 1881 pairs in the NKR data set, this simulation gives an average waiting time and an empirical probability of being matched under both the greedy and patient policies. Figure 9(a) shows that for each pair, the average waiting time is shorter under the greedy policy than under the patient policy; all of the dots are above the 45° line. This observation suggests that the waiting time distribution under the greedy policy first-order stochastically dominates the waiting time distribution under the patient policy.[31] Figure 9(b) reports the match rates under the greedy and patient policies. Observe that for most pairs the empirical probabilities of matching under the greedy and patient policies are "close" to the 45 degree line suggesting that the probability of being matched is roughly the same for *every pair*. Interestingly, Figure 9(b) suggests that under the greedy policy, easy-to-match pairs are slightly worse off because they are matched with slightly lower probability, whereas hard-to-match pairs are better off.

[28] An X–Y patient–donor pair contains a patient with bloodtype X and a donor with bloodtype Y.

[29] Observe that A–O pairs (which are over-demanded) can match potentially match with each other and with O–A pairs (under-demanded) but O–A pairs cannot match with each other.

[30] More than 40% of patients in over-demanded pairs have less than a 5% chance of being tissue-type compatible with a random donor. Furthermore, about 10% of over-demanded pairs have no match within this data set, which is why the average waiting times do not drop all the way to zero.

[31] A detailed analysis of the simulation results confirms that this is indeed the case. We omit the details.



Waiting times

Chance of matching

FIGURE 9

Averages of waiting times (left) and chance of matching (right) taken over copies for each pair in the data. The axes correspond to the greedy and patient policies.

6. A DETAILED DISCUSSION OF OUR MODELLING ASSUMPTIONS

From a modelling perspective, there are three major differences between our paper and the closely related literature on dynamic matching (Ünver, 2010; Ashlagi, Jaillet and Manshadi, 2013; Anderson *et al.*, 2017; Akbarpour, Combe, He, Hiller, Shimer and Tercieux, 2019; Akbarpour *et al.*, 2020; Nikzad, Akbarpour, Rees and Roth, 2019):

(i) Compatibilities in our model depend on the agents' types (blood types) and a random component (sensitization of the patient).[32]

(ii) We focus on markets where the compatibility probabilities do not vanish with the market size.

(iii) The objective of expected waiting time and probability of being matched differs from the objectives considered in some of the literature.

The two-type compatibility model. One may interpret our two-type model as a stylized way of capturing both blood types and randomness due to tissue-type incompatibilities. Intuitively, hard-to-match agents in our model correspond to pairs who cannot match with each other due to blood type incompatibility (for instance O–A blood type patient–donor pairs).[33] Due to the presence of such agents in kidney exchanges, not all agents can be matched even when the market grows large. As we have argued in Section 2.1, no single-type model can reproduce this feature of real kidney exchanges while at the same time providing each pair at least one potential match. Furthermore, one motivation for our article is the concern, sometimes raised by practitioners, that greedy matching might harm hard-to-match agents. This concern can by definition not be addressed in a single-type model where all agents are equally difficult to match.

[32] Anderson *et al.* (2017); Akbarpour *et al.* (2020) are examples of papers that consider a single-type model. A notable exception is Ünver (2010) who analyses matching policies in a deterministic compatibility model with multiple compatibility types, but without agent criticality times or random compatibility.

[33] Similar type of asymmetries across agents also appear in Nikzad *et al.* (2019). They are concerned with a proposal for global kidney exchange, which incorporates international pairs to domestic kidney exchange pools. Their model takes a reduced form approach where there is a continuum of international pairs who do not get matched to each other and a continuum of domestic pairs who can get matched to each other and to the international pairs. The compatibilities (between measures of pairs) are determined by a "matching function". They do a steady-state analysis to answer whether the savings from dialysis can cover the surgery costs of international pairs.

As a robustness exercise we also ran simulations for our model when there is only a single type, i.e., $\lambda = -1$. We report the result in Section vi in the Supplementary appendix. These simulations indicate that the greedy policy remains optimal in large markets even with a single type.

Non-vanishing compatibility. Previous literature (Ashlagi *et al.*, 2013; Anderson *et al.*, 2017; Akbarpour *et al.*, 2019, 2020; Nikzad *et al.*, 2019) considered models with vanishing compatibility probabilities, i.e., when the arrival rate of agents equals m, the probability of compatibility of some pairs of agents equals $\frac{c}{m}$. In contrast, the probability of two patient–donor pairs being compatible in our model is independent of the market size. Assuming that the compatibility probability depends on the market size is intended to capture small and sparse kidney exchanges. In contrast, assuming that this probability is unaffected by the market size is natural when considering large kidney exchanges. Whether a given market is approximated by either compatibility model depends on the specific context and is ultimately an empirical question. As we explained in Section 2.1, the combination of having two types and non-vanishing compatibilities allows us to capture crucial features of the compatibility graph in kidney exchanges.

We provide simulations for our model with vanishing compatibility probabilities in the Supplementary Appendix. These simulations indicate if the compatibility probability approaches zero at rate $\frac{1}{\sqrt{m}}$, greedy is optimal in a large market while the patient and batching policies with a fixed batch length are not (Section vi in the Supplementary appendix). The simulations also show that this is not necessarily true if the compatibility probability vanishes at rate $\frac{1}{m}$. In this case, there is a trade-off and the greedy policy leads to a lower waiting time while the patient policy matches more agents.

Different objectives. Another difference with some of the literature is the objective we consider. Akbarpour *et al.* (2020) consider the *loss rate*, denoted by $L_\pi \in [0,1]$ for a policy π, which is the probability that an agent is *not* matched under the policy π. To compare two policies π, π', they consider the *ratio of loss rates* $L_\pi / L_{\pi'}$. We focus on the probability of being matched in a policy and expected waiting time. One reason for studying match rate and waiting time is that they together determine the payoff of a risk-neutral EU who assigns a fixed value to being matched and incurs a constant cost while waiting in the market, whereas the ratio between the loss rates is not related to EU preferences. For example, we can have two policies π, π' which both match almost everyone (and thus achieve a loss rate close to zero) such that the ratio of loss rates $L_\pi / L_{\pi'}$ is arbitrarily large yet policy π matches agents much faster than π'. In the limit, every risk-neutral EU maximizer prefers π over π' even though the policy π' is arbitrarily much better according to the ratio of loss rate.[34] In our model, the loss rates will converge to the same non-zero limit in a large market for the greedy and patient policy (as we established in Proposition 1). Thus, the greedy and patient policies are not ordered according to the ratio of loss rates, but any decision maker who considers a combination of the loss rates and the waiting time would prefer the greedy over the patient policy in a sufficiently large market. In the kidney exchange context both policies lead to almost the same probability of receiving a match, whereas pairs match much faster under the greedy policy than under the patient policy (see our experiments in Section 5).

[34] Consider an EU agent, with utility $u_\pi = vp - ct$ under policy π, whose value for being matched is $v > 0$ and her waiting cost per unit of time is $c > 0$. Consider two policies π, π', with match rates p, p' and waiting times t, t'. Suppose $p = 1 - 10^{-k}, p' = 1 - 10^{-2k}$, and $t = t'$, where k is a sufficiently large positive number. Then, $L_\pi / L_{\pi'} = 10^k$, whereas $u_\pi - u_{\pi'} = (10^{-k} - 10^{-2k})v$ approaches 0 as k grows large. Furthermore, when tâŁ™$- t > \frac{(10^{-k} - 10^{-2k})v}{c}$, it holds that $u_{\pi'} < u_\pi$. Hence, even though there is large gap between the loss ratios of the policies (in favour of π'), the agent's expected utility would be larger under π.

7. CONCLUSION

This article studies matching policies in a random dynamic market in which some agents are easier to match than others. We show theoretically as well as empirically that the greedy matching policy is arbitrarily close to optimal for all agents in sufficiently large markets. This finding has direct practical implications for kidney exchanges that may not employ greedy matching policies out of concern that greedy matching may harm those patients for whom it is hardest to find a compatible partner (Ferrari *et al.*, 2014). Our simulations further suggest that matching frequently does not reduce the number of transplants even for realistic market sizes.[35]

This article has some limitations in the context of kidney exchange. We only considered pairwise matchings and ignored frictions that occur in practice. Simulations in Ashlagi *et al.* (2018) account for such frictions, three-way cycles, and chains.[36] They find that the greedy policy is optimal among a class of batching policies. We conjecture that this holds also true within our model, and that the benefit of matching in chains or longer cycles vanishes in a sufficiently large markets. It remains an interesting question to study these questions theoretically in small markets.

Throughout, we abstracted away from match qualities and assumed that agents are indifferent to whom they match with. When preferences over match partners play an important role, the greedy policy might not be optimal. For example, if matching easy-to-match agents with each other creates a much larger value than matching an easy-to-match agent to a hard-to-match one, the greedy policy can be suboptimal since almost all easy-to-match agents will match with hard-to-match agents in a large market (for studies with match qualities, see Baccara *et al.*, 2020; Li, Lieberman, Macke, Carrillo, Ho, Wellen and Das, 2019; Mertikopoulos *et al.*, 2020; Blanchet *et al.*, 2020; Aquilina, Budish and O'Neill, 2020).

We also abstracted away from the incentives agents might have to misreport their private information, such as arrival time, type, and realized compatibilities. One may ask what policy would be optimal if any of these would be the agent's private information. Under greedy the agent would have no incentive to delay reporting her arrival or claiming to be incompatible with agents whom she is compatible with; doing so would lead her to be matched later and less often. However, easy-to-match agents, after some histories, may prefer to misreport their type since our greedy policy break ties in favour of hard-to-match agents. To satisfy incentive compatibility, one would need to break ties uniformly, which does not affect the asymptotic optimality of the greedy policy. Notably, this conclusion holds since agents are indifferent between who they match with.[37]

Acknowledgments. We want to thank Mohammad Akbarpour, Aaron Bodoh-Creed, Yeon-Koo Che, Yuichiro Kamada, and Olivier Tercieux for useful comments and suggestions. Ashlagi acknowledges the research support of the National Science Foundation grant (SES-1254768). Philipp Strack was supported by a Sloan fellowship.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

[35] Independent of the policy, increasing the market size by merging pools can improve the match rate and waiting times of agents. So, competition between different kidney exchanges can harm the number of transplants but not because of frequent matching.

[36] For many exchanges, chains are not a practical concern as numerous programs have very low enrolment of altruistic donors that initiate chains and, in some countries like France, Poland, and Portugal, chains are not even feasible since altruistic donation is illegal (Biro *et al.*, 2017).

[37] Baccara *et al.* (2020) study incentive compatible matching policies in a setting with match qualities.

APPENDIX

A. PROOFS FOR PROPOSITIONS 1 AND 2

Proof of Proposition 1 Note that, as there are more E agents than H agents and H agents cannot match to themselves, $\frac{2}{2+\lambda}$ is an upper bound on the fraction of agents which can be matched for any m. Note, that the size of the maximal matching (SMM) equals $\frac{2}{2+\lambda}$ if the bipartite graph with m easy-to-match agents and m hard-to-match agents on the other side admits a perfect matching. The probability that such a perfect matching exists converges to one as $m\to\infty$ (see e.g. Theorem 5.1 page 77 in Frieze and Karoński (2015)). This proves the claim about SSM.

The probability that a hard-to-match agent has no partner is given by $(1-p)^m$. Because the compatibilities between hard-to-match and easy-to-match agents are drawn independently, the probability that all hard-to-match agents have at least one partner is given by $(1-(1-p)^m)^{m(1+\lambda)}$. This probability converges to one as $m\to\infty$. The same argument shows that the probability that all easy-to-match agents have at least one partner approaches one as m approaches infinity. □

Proof of Proposition 2 The proof is by contradiction; suppose such $p(m)$ exists. The chance that an agent has no other compatible agents is $(1-p(m))^m$. If $p(m)=O(1/m)$, then for sufficiently large m we have

$$(1-p(m))^m > e^{-2mp(m)} = e^{-O(1)},$$

since $1-\alpha > e^{-2\alpha}$ for $\alpha\in(0,\frac{1}{2})$. Thus, (2) cannot be satisfied. Therefore, suppose that $p(m)=\frac{\omega(m)}{m}$, where $\lim_{m\to\infty}\omega(m)=\infty$. Next, we use this property to show that (1) cannot be satisfied.

The proof is constructive. We propose a simple algorithm that chooses a matching μ with size $|\mu|$ such that $\lim_{m\to\infty}\frac{|\mu|}{m}=1$. Our algorithm is a greedy algorithm, defined as follows. It orders agents of the graph from 1 to m and visits the agents one by one. When visiting agent i, if there are no agents left that are compatible with agent i, then the algorithm passes agent i and moves to agent $i+1$. Otherwise, the algorithm chooses one of the neighbours of agent i arbitrarily, namely agent j, and adds the pair (i,j) to the matching. The algorithm then visits the next available agent in the ordering. This process continues until the algorithm visits all agents.

We claim that the algorithm produces a matching μ which satisfies $\lim_{m\to\infty}\frac{|\mu|}{m}=1$. Let $\phi(m)$ be a function that grows faster than $\frac{m}{w(m)}$ but slower than m. Then, during the algorithm, so long as there are $\phi(m)$ unvisited agents, the chance that a visited agent has no compatible agents is

$$(1-w(m)/m)^{\phi(m)} \leq e^{-\frac{w(m)\phi(m)}{m}} = o(1).$$

Hence, so long as there are $\phi(m)$ agents left in the graph, the agent visited by the algorithm will be matched with a probability at least $1-q(m)$ where $\lim_{m\to\infty}q(m)=0$. By linearity of expectation, the expected number of unmatched agents by the end of the algorithm is then at most $\phi(m)+(m-\phi(m))\cdot q(m)$. Noting that $\lim_{m\to\infty}\frac{\phi(m)+(m-\phi(m))\cdot q(m)}{m}=0$ completes the proof. □

B. PRELIMINARIES

Let m_Θ denote the arrival rate of agents of type $\Theta\in\{E,H\}$. We use the terms *E pool* and *H pool* to denote the pools containing only E agents and only H agents, respectively. The *criticality clock* of an agent refers to the exponential random variable that determines the exogenous time that an agent becomes critical. Immediately after the criticality clock of an agent present in the pool *ticks*, she departs the pool. (Throughout the draft, the term *departure* is used to refer to the event of an agent leaving the pool, either matched or unmatched.)

We next describe how the greedy and patient policies break ties between feasible matches. Consider an agent, say a, who arrives to the market under the greedy policy or gets critical under the patient policy. At the time of this event, both policies attempt to match a as follows. First, a strict order over all H agents in the market is selected uniformly at random, and a is matched with the first compatible H agent according to the selected order. If such an H agent does not exist, then a strict order over all E agents in the pool is selected uniformly at random, and a is matched with the first compatible E agent in that order if such an agent exists.

B.1. *Asymptotic notions*

We say a statement $\mathcal{S}(i)$ holds for *sufficiently large i* if there exists i_0 such that $\mathcal{S}(i)$ holds for all $i>i_0$. Let $E(i)$ be an event parameterized by a positive integer i. We say that $E(i)$ occurs with *high probability as i grows large* if $\lim_{i\to\infty}\mathbb{P}[E(i)]=1$. We often let the parameter i be m, the arrival rate of easy-to-match agents. When this is clearly known from the context, we simply say that $E(m)$ occurs *with high probability* or, briefly, $E(m)$ occurs *whp*.

Furthermore, we say that $E(i)$ occurs *with very high probability as i grows large* if there exists $\alpha>1$ such that $\lim_{i\to\infty}\frac{1-\mathbb{P}[E(i)]}{e^{-(\ln i)^\alpha}}=0$. We often let the parameter i be m, the arrival rate of easy-to-match agents. When this is clearly known from the context, we simply say that $E(m)$ occurs *with very high probability* or, briefly, $E(m)$ occurs *wvhp*.

For any two functions $f,g:\mathbb{R}_+\to\mathbb{R}_+$ we adopt the notation $f=o(g)$ when for every positive constant ϵ there exists a constant i_ϵ such that $f(i)\leq\epsilon g(i)$ holds for all $i>i_\epsilon$. We define $f=O(g)$ if there exist positive constants i_0,Δ such that $f(i)\leq g(i)\Delta$ holds for all $i>i_0$.

B.2. *Markov chains*

We denote the state space of a Markov chain $\mathcal{X}$ by $V(\mathcal{X})$.

Proposition 9 (Anderson *et al.*, 2017, Proposition EC.4) *Let $\mathcal{X}=(X_0,X_1,X_2,X_3,\ldots)$ be a discrete time positive recurrent Markov chain with a countable state space and steady-state distribution η. Also, let $\mathbb{E}_x[\cdot]$ denote the expectation operator conditional on $X_0=x$. Suppose that there exist real numbers $\alpha,\beta\geq 0$ and $\gamma>0$, a set $B\subset S$, and functions $U:V(\mathcal{X})\to\mathbb{R}_+$ and $f:V(\mathcal{X})\to\mathbb{R}_+$ such that for $x\in V(\mathcal{X})\backslash B$,*

$$\mathbb{E}_x[U(X_1)-U(X_0)]\leq-\gamma f(x), \tag{B.1}$$

and for $x\in B$,

$$f(x)\leq\alpha, \tag{B.2}$$

$$\mathbb{E}_x[U(X_1)-U(X_0)]\leq\beta. \tag{B.3}$$

Then,

$$\mathbb{E}_{X\sim\eta}\big[f(X)\big]\leq\alpha+\frac{\beta}{\gamma}. \tag{B.4}$$

The stochastic processes associated with our matching policies are continuous-time processes. The above proposition, however, is applicable to discrete-time processes. To close this gap, we use the notion of *embedded Markov chain*.

Embedded Markov chain Let $\mathcal{X}$ be a continuous-time Markov chain with a countable state space. For any two states of $\mathcal{X}$, namely i,j, let $n_{i,j}$ denote the transition rate from state i to state j. Let N be the *transition rate matrix* for $\mathcal{X}$, i.e., $N_{i,j}=n_{i,j}$ for $i\neq j$, and the entries on the diagonal of N are set so that each row in N sums to 0.

Definition 5. *The embedded Markov chain of $\mathcal{X}$, denoted by $\widehat{\mathcal{X}}$, is a discrete-time Markov chain with the same state space as $\mathcal{X}$. The transition probability from state i to state j in $\widehat{\mathcal{X}}$ is denoted by $\widehat{n}_{i,j}$ and is defined by*

$$\widehat{n}_{i,j}=\begin{cases}\frac{n_{i,j}}{\sum_{k\neq i}n_{i,k}} & \textit{if } i\neq j\\ 0 & \textit{if } i=j.\end{cases}$$

Fact 2 (Harchol-Balter (2013)) *Let $\mathcal{X}$ be an ergodic continuous-time Markov chain with a unique stationary distribution ρ and transition rate matrix N. Then, the embedded Markov chain of $\mathcal{X}$, namely $\widehat{\mathcal{X}}$ has a unique steady-state distribution, namely $\widehat{\rho}$. Furthermore, for every state $i\in V(\mathcal{X})$,*

$$\rho(i)=\frac{\widehat{\rho}(i)/N_{i,i}}{\sum_{j\in V(\mathcal{N})}\widehat{\rho}(j)/N_{j,j}}.$$

B.3. *Inequalities*

Fact 3 (Canonne, 2019) *For a Poisson random variable X with mean μ, it holds that*

$$\mathbb{P}[|X-\mu|>z]\leq 2e^{-\frac{z^2}{\mu+z}}. \tag{B.5}$$

Chernoff bounds are concentration inequalities that bound the deviations of a weighted sum of Bernouli random variables from its mean. Below we present their multiplicative form.

Fact 4 (Chernoff bound Brémaud 2017) *Let $X_1,\ldots,X_n$ be a sequence of n independent random binary variables such that $X_i=1$ with probability p_i and $X_i=0$ with probability $1-p_i$. Let $\alpha_1,\ldots,\alpha_n$ be arbitrary real numbers in the unit interval. Also, let $S=\sum_{i=1}^n\alpha_i\mathbb{E}[X_i]$. Then, for any ϵ with $0\leq\epsilon\leq 1$, we have:*

$$\Pr\big[\textstyle\sum_{i=1}^n\alpha_iX_i>(1+\epsilon)S\big]\leq e^{-\epsilon^2S/3},$$

$$\Pr\big[\textstyle\sum_{i=1}^n\alpha_iX_i<(1-\epsilon)S\big]\leq e^{-\epsilon^2S/2}.$$

Fact 5. *For reals A,B, at every $x \neq -B$, the function $g(x)=\frac{x+A}{x+B}$ is increasing in x iff $A \leq B$.*

Proof. We observe that $g'(x)=\frac{B-A}{(B+x)^2}$. Hence, when $x \neq -B$, $g'(x) \geq 0$ holds iff $A \leq B$. □

Fact 6. *For every real A,B, at every $x \neq -B/2$, the function $g(x)=\frac{x-A}{2x+B}$ is increasing in x.*

Proof. Observing that $g'(x)=\frac{2A+B}{(B+2x)^2}$ proves the claim. □

C. PROOF OF PROPOSITION 4

We fix $\lambda > 0$ throughout the proof.

Definition 6. *For a policy τ and every agent type $\Theta \in \{E,H\}$, let $q^{\tau}_{\Theta}(m)$ and $w^{\tau}_{\Theta}(m)$, respectively denote the match rate and the expected waiting time of agents of type Θ under the policy τ when the arrival rates of E and H agents are respectively m and $m(1+\lambda)$.*

Lemma 1. *For every agent type $\Theta \in \{E,H\}$, $q^{\tau}_{\Theta}(m) \geq 1-\frac{w^{\tau}_{\Theta}(m)}{d}$.*

Proof. Let $\mathbb{E}_t[\cdot], \mathbb{P}_t[\cdot]$ denote the expectation and probability operator conditional on all information the planner has at time t. Consider an agent i with an arbitrary type $\Theta \in \{E,H\}$. Recall that φ_i is the random variable denoting the difference between the time an agent i arrives to the market and the time that she departs (whether matched or not). Let κ_i be the difference between the time an agent i arrives to the market and the time she becomes critical. By definition, κ_i is exponentially distributed with mean d. Since the policy does not observe the value of κ_i before the agent becomes critical, then

$$\mathbb{P}_{\alpha_i}[\varphi_i < t \,|\, \kappa_i = t] = \mathbb{P}_{\alpha_i}[\varphi_i < t \,|\, \kappa_i \geq t].$$

This implies that the probability that the agent departs strictly before she becomes critical is given by

$$\begin{aligned}
\mathbb{P}_{\alpha_i}[\varphi_i < \kappa_i] &= \int_0^{\infty} \mathbb{P}_{\alpha_i}[\varphi_i < \kappa_i \,|\, \kappa_i = t]\frac{1}{d}e^{-\frac{1}{d}t}dt = \int_0^{\infty} \mathbb{P}_{\alpha_i}[\varphi_i < t \,|\, \kappa_i \geq t]\frac{1}{d}e^{-\frac{1}{d}t}dt \\
&= \int_0^{\infty} (1-\mathbb{P}_{\alpha_i}[\varphi_i \geq t \,|\, \kappa_i \geq t])\frac{1}{d}e^{-\frac{1}{d}t}dt = 1-\int_0^{\infty} \mathbb{P}_{\alpha_i}[\varphi_i \geq t \,|\, \kappa_i \geq t]\mathbb{P}_{\alpha_i}[\kappa_i \geq t]\frac{1}{d}dt \\
&= 1-\int_0^{\infty} \mathbb{P}_{\alpha_i}[\varphi_i \geq t]\frac{1}{d}dt = 1-\frac{1}{d}\mathbb{E}[\varphi_i].
\end{aligned}$$

As the agent always departs the market matched whenever she departs before becoming critical, then a lower bound on the probability that this agent is matched is given by $1-\frac{1}{d}\mathbb{E}_{\alpha_i}[\varphi_i]$, i.e., $1-\frac{1}{d}\mathbb{E}_{\alpha_i}[\varphi_i] \leq \mathbb{E}_{\alpha_i}[\mu_i]$. Taking the average over agents of type Θ and using the law of iterated expectations yields that

$$\begin{aligned}
1-\frac{w^{\tau}_{\Theta}(m)}{d} &= 1-\frac{1}{d}\lim_{t\to\infty}\mathbb{E}\left[\frac{\sum_{i:\,\alpha_i \leq t \text{ and } \theta_i=\Theta}\varphi_i}{|\{i:\,\alpha_i \leq t \text{ and } \theta_i=\Theta\}|}\right] = \lim_{t\to\infty}\mathbb{E}\left[\frac{\sum_{i:\,\alpha_i \leq t \text{ and } \theta_i=\Theta} 1-\frac{1}{d}\varphi_i}{|\{i:\,\alpha_i \leq t \text{ and } \theta_i=\Theta\}|}\right] \\
&= \lim_{t\to\infty}\mathbb{E}\left[\frac{\sum_{i:\,\alpha_i \leq t \text{ and } \theta_i=\Theta} 1-\frac{1}{d}\mathbb{E}_{\alpha_i}[\varphi_i]}{|\{i:\,\alpha_i \leq t \text{ and } \theta_i=\Theta\}|}\right] \leq \lim_{t\to\infty}\mathbb{E}\left[\frac{\sum_{i:\,\alpha_i \leq t \text{ and } \theta_i=\Theta}\mathbb{E}_{\alpha_i}[\mu_i]}{|\{i:\,\alpha_i \leq t\}|}\right] \\
&= \lim_{t\to\infty}\mathbb{E}\left[\frac{\sum_{i:\,\alpha_i \leq t \text{ and } \theta_i=\Theta}\mu_i}{|\{i:\,\alpha_i \leq t\}|}\right] = q^{\tau}_{\Theta}(m).
\end{aligned}$$

□

Proof of Proposition 4 Consider a policy τ, and let A_t and B_t, respectively denote the number of hard-to-match and easy-to-match agents that arrive prior to time t under the policy τ. Since hard-to-match agents can be matched only to easy-to-match agents, the match rate of hard-to-match agents is at most $\lim_{t\to\infty}\frac{B_t}{A_t}$, by the Ergodic theorem. By the strong law of large numbers, $\lim_{t\to\infty}\frac{B_t/t}{A_t/t}=\frac{1}{1+\lambda}$. This proves that $q^{\tau}_{H}(m) \leq \frac{1}{1+\lambda}$ and establishes the claim about the match rate.

To prove the result for waiting time, we use Lemma 1 to write $q^{\tau}_{H}(m) \geq 1-\frac{w^{\tau}_{H}(m)}{d}$. On the other hand, we showed that $\frac{1}{1+\lambda} \geq q^{\tau}_{H}(m)$. Hence, $\frac{1}{1+\lambda} \geq 1-\frac{w^{\tau}_{H}(m)}{d}$, which means that $w^{\tau}_{H}(m) \geq \frac{d\lambda}{1+\lambda}$. □

D. GREEDY MATCHING: CONCENTRATION BOUND

Here, we analyse the stochastic process corresponding to the greedy policy and develop a concentration bound on the number of hard-to-match agents present in the pool. Because one can renormalize the time scale and the arrival rates linearly with a factor of $1/d$, we suppose that $d=1$ throughout this section. This is without generality, as speeding up or slowing down time does not change the steady-state distribution of the number of hard- or easy-to-match agents in the pool. We further assume that $m \geq 1$ throughout this section, i.e., at least one agent arrives per year.

The main technical results established by this analysis are a lower and an upper concentration bound for the number of hard-to-match agents in the greedy policy, developed respectively in Sections D.2 and D.3. Using these concentration bounds, we will be able to prove the results on the match rate and waiting time under the greedy policy, which appear in Section E.

D.1. *Modelling the dynamics*

We use a two-dimensional continuous-time Markov chain, which we denote by $\mathcal{M}$, to model the dynamics of the market. First, we set up some notation before proceeding to the description. For a Markov chain $\mathcal{M}$, we recall that $V(\mathcal{M})$ denotes the state space of $\mathcal{M}$. We represent each state by a pair (x,y) where x,y respectively denote the number of H agents and the number of E agents. In other words, we have

$$V(\mathcal{M})=\{(x,y):\ x,y\in\mathbb{Z} \text{ and } x,y\geq 0\}.$$

By definition, the Markov chain is at state (x,y) if there are x hard-to-match and y easy-to-match agents in the pool.

Lemma 2. *$\mathcal{M}$ has a unique stationary distribution.*

The above lemma is proved in the Supplementary Appendix, Section i. The proof uses standard arguments that bound the expected return time to a fixed state.

Definition 7. *Let π denote the stationary distribution of $\mathcal{M}$, with $\pi_{x,y}$ denoting the probability that π assigns to a state (x,y), and let $\pi_x=\sum_{y=0}^{\infty}\pi_{x,y}$ be the associated marginal distribution of x.*

D.2. *A large market lower concentration bound for greedy matching*

In this section, we provide the core technical result for the analysis of greedy matching: Theorem 3. This theorem provides a concentration bound for the stochastic process associated with the greedy policy. We state and prove Theorem 3 in Section D.2.4, after the following preliminary analysis.

We first define and analyse a new stochastic process called *simplified greedy* which differs from greedy in that E agents are not matched with each other. We prove that the number of H agents present in the market in the new process is lower (in first-order stochastic dominance) than the number of H agents present under the greedy policy, and use this to prove the concentration bound.

D.2.1. The simplified greedy process and its associated Markov chain.

Definition 8. *The simplified greedy process is the same as the greedy process with the difference that, in the simplified greedy process, E agents are considered to be incompatible (and therefore are never matched to each other).*

We denote by $\mathcal{N}$ the two-dimensional, continuous-time Markov chain counting the number of H and E agents present at every point in time under the simplified greedy policy. As before, we denote by x the number of H agents and by y the number of E agents. We next describe the transition rates of $\mathcal{N}$. A transition can only happen from a state (x,y) to its (at most) four *neighbours*,

$$\{(x',y')\in\mathbb{Z}_+^2:\ |x-x'|+|y-y'|=1\}.$$

See Figure D.1 for a visual depiction of the neighbours and transition rates. To simplify the definition of transition rates from a node to its neighbours, we define the following notations: Let $N_x=(1-p)^x$ and $\bar{N}_x=1-N_x$. (Thus, N_x is the probability that an E agent is incompatible with x H agents.) For each state (x,y), we denote the transition rates from this state to its neighbour on the top, right, bottom, and left by $u_{x,y},r_{x,y},d_{x,y},l_{x,y}$, respectively. These rates are defined as follows:

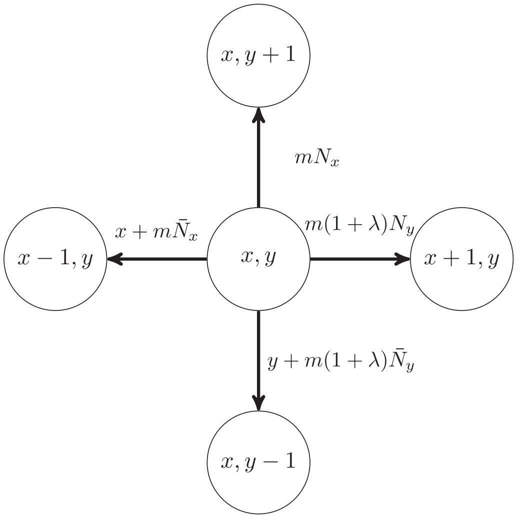


FIGURE D.1
An illustration of the transitions from node (x,y) to its neighbours.

- $u_{x,y}=mN_x$ is the transition rate from the node (x,y) to node $(x,y+1)$. This holds because E agents arrive with rate m; after the arrival of an E agent, the number of E agents increases by one if the arriving E agent is not compatible to any H agent present in the pool.
- $r_{x,y}=m(1+\lambda)N_y$ is the transition rate from the node (x,y) to node $(x+1,y)$. This holds because H agents arrive with rate $(1+\lambda)m$; after the arrival of an H agent, the number of H agents increases by one if the arriving H agent is incompatible to all E agents in the pool.
- $d_{x,y}=y+m(1+\lambda)\bar{N}_y$ is the transition rate from the node (x,y) to node $(x,y-1)$. This holds because the number of E agents goes down by one when (i) a new H agent arrives who is compatible to an E agent; this happens with rate $m(1+\lambda)\bar{N}_y$; (ii) an existing E agent becomes critical and departs the pool; this happens with rate y.
- $l_{x,y}=x+m\bar{N}_x$ is the transition rate from the node (x,y) to node $(x-1,y)$. This holds because the number of H agents goes down by one when (i) a new E agent arrives who is compatible to an H agent; this happens with rate $m\bar{N}_x$; (ii) an existing H agent becomes critical and departs the pool; this happens with rate x.

Lemma 3. *$\mathcal{N}$ has a unique stationary distribution.*

Proof. The proof is identical to the proof of Lemma 2, but for the letter $\mathcal{M}$ replaced with $\mathcal{N}$. □

Definition 9. *Let ρ denote the stationary distribution of $\mathcal{N}$, with $\rho_{x,y}$ denoting the probability that ρ assigns to the state (x,y). Also, let $\rho_x=\sum_{y=0}^{\infty}\rho_{x,y}$.*

Next we show that, at the steady state, fewer hard-to-match agents wait in the simplified greedy process $\mathcal{N}$ than in the original greedy process $\mathcal{M}$ in the sense of first-order stochastic dominance.

Lemma 4. *For every $x\geq 0$, $\sum_{i=0}^{x}\pi_i\leq\sum_{i=0}^{x}\rho_i$.*

The proof of the lemma is technical and is deferred to the Supplementary Appendix, Section i. The proof idea is defining a *coupling* of $\mathcal{M}$ and $\mathcal{N}$ such that, in the coupled process, there are more H agents in the pool under $\mathcal{M}$ than under $\mathcal{N}$ at any time, and fewer E agents.

Definition 10. *Let $\widehat{\mathcal{N}}$ denote the embedded Markov chain corresponding to $\mathcal{N}$. Also, let $\hat{\rho}$ denote its unique stationary distribution.*

In the above definition, we recall that $\hat{\rho}$ exists and is unique by Fact 2.

D.2.2. Applying Proposition 9 to $\widehat{\mathcal{N}}$. We develop a concentration bound for $\hat{\rho}$ using Proposition 9. This bound is parameterized by a number $k>0$. We let $k>0$ be an arbitrary real number in the following analysis.

Let $x^*=\lambda m$. Define $B_k=\{(x,y)\in\mathbb{Z}_+^2 : |x-x^*|+|y|<k\sqrt{m}\}$. Let the functions f,U be:

$$f(x,y)=\mathbb{1}_{(x,y)\notin B_k},\ U(x,y)=e^{\frac{y+|x-x^*|}{\sqrt{m}}},$$

where $\mathbb{1}_{(x,y)\notin B_k}$ is the indicator function that equals 1 if $(x,y)\notin B_k$. We will apply Proposition 9 to $\widehat{\mathcal{N}}$. To this end, we need the following definitions.

Definition 11. *Conditional on the Markov chain $\widehat{\mathcal{N}}$ being at a state (x,y), let the random variable (x_1,y_1) denote the next state that the Markov chain moves to. Define*

$$\Delta(x,y)=\mathbb{E}\big[U(x_1,y_1)\big]-U(x,y).$$

Definition 12. *For every $z\geq 0$, define*

$$H(z)=-e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{z-2m(1+\lambda)N_z}{2m(1+\lambda)+z}-\frac{1}{m}\right).$$

Lemma 5. *For every $x>x^*$ and $y\geq 0$, $\Delta(x,y)\leq H(y+x-x^*)$.*

Proof. To shorten notation, let $n=(1+\lambda)m$, $\theta=m+n+x+y$, and $z=y+x-x^*$. Note that $z=y+|x-x^*|$. Using this notation, we can write $U(x,y)=e^{\frac{z}{\sqrt{m}}}$. Recall that we defined $N_\alpha=(1-p)^\alpha$ and $\bar{N}_\alpha=1-N_\alpha$ for every real $\alpha>0$. Then,

$$\Delta(x,y)=e^{\frac{z+1}{\sqrt{m}}}\left(\frac{mN_x+nN_y}{\theta}\right)+e^{\frac{z-1}{\sqrt{m}}}\left(\frac{x+m\bar{N}_x+y+n\bar{N}_y}{\theta}\right)-e^{\frac{z}{\sqrt{m}}} \tag{D.6}$$

$$=e^{\frac{z}{\sqrt{m}}}\left(e^{\frac{1}{\sqrt{m}}}\frac{mN_x+nN_y}{\theta}+e^{-\frac{1}{\sqrt{m}}}\left(\frac{x+m\bar{N}_x+y+n\bar{N}_y}{\theta}\right)-1\right)$$

$$\leq e^{\frac{z}{\sqrt{m}}}\left(\left(1+\frac{1}{\sqrt{m}}+\frac{1}{m}\right)\frac{mN_x+nN_y}{\theta}+\left(1-\frac{1}{\sqrt{m}}+\frac{1}{m}\right)\left(\frac{x+m\bar{N}_x+y+n\bar{N}_y}{\theta}\right)-1\right) \tag{D.7}$$

$$=e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2mN_x+2nN_y-m-n-x-y}{\theta}+\frac{1}{m}\right) \tag{D.8}$$

$$=e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2mN_x+2nN_y-\theta}{\theta}+\frac{1}{m}\right). \tag{D.9}$$

To see why (D.6) holds, observe that $\frac{mN_x}{\theta}$ and $\frac{nN_y}{\theta}$ are the transition probabilities of $\widehat{\mathcal{N}}$ from the state (x,y) to the states $(x,y+1)$ and $(x+1,y)$, respectively. When either of these transitions occur, the value of U changes from $e^{\frac{z}{\sqrt{m}}}$ to $e^{\frac{z+1}{\sqrt{m}}}$. Also, $\frac{x+m\bar{N}_x}{\theta}$ and $\frac{y+n\bar{N}_y}{\theta}$ are the transition probabilities of $\widehat{\mathcal{N}}$ from the state (x,y) to the states $(x-1,y)$ and $(x,y-1)$, respectively. Note that these transition probabilities are 0 if $x-1$ or $y-1$ are negative. When either of these transitions occur, the value of U changes from $e^{\frac{z}{\sqrt{m}}}$ to $e^{\frac{z-1}{\sqrt{m}}}$. Inequality (D.7) holds because $e^\alpha\leq 1+\alpha+\alpha^2$ holds for every $\alpha\in[-1,1]$. Equations (D.8) and (D.9) hold by rearrangement of terms.

Claim 1. *Let a,b,c be positive reals such that $a<b$. The function $g(s)=(1-p)^{b-s}+c(1-p)^{s-a}$ is convex over $[a,b]$.*

Proof. Observe that

$$g''(s)=(1-p)^{-s}\log^2(1-p)\Big(c(1-p)^{2s-a}+(1-p)^b\Big)\geq 0,$$

which means that g is convex when $p\in(0,1)$. Also, when $p=1$, $g(s)=0$ for every $s\in[a,b]$. □

Claim 2.

$$e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2m(N_{\lambda m+z}+1+\lambda)-\theta}{\theta}+\frac{1}{m}\right)\leq H(z).$$

Proof. Recall that $\theta=m+n+x+y$, which means that $\theta=2m(1+\lambda)+z$. Hence, to prove the claim, which says that

$$e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2m(N_{\lambda m+z}+1+\lambda)-\theta}{\theta}+\frac{1}{m}\right)\leq e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2m(1+\lambda)N_z-z}{2m(1+\lambda)+z}+\frac{1}{m}\right),$$

it suffices to prove that

$$2m(N_{\lambda m+z}+1+\lambda)-\theta \leq 2m(1+\lambda)N_z-z,$$

or equivalently, $2mN_{\lambda m+z}-z \leq 2m(1+\lambda)N_z-z$. The latter inequality holds as $N_{\lambda m+z} \leq N_z$. □

Claim 3.

$$e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2m(N_{\lambda m}+(1+\lambda)N_z)-\theta}{\theta}+\frac{1}{m}\right) \leq H(z)$$

Proof. Recall that $\theta = 2m(1+\lambda)+z$. Hence, to prove that

$$e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2m(N_{\lambda m}+(1+\lambda)N_z)-\theta}{\theta}+\frac{1}{m}\right) \leq e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2m(1+\lambda)N_z-z}{2m(1+\lambda)+z}+\frac{1}{m}\right),$$

it suffices to prove that

$$2m(N_{\lambda m}+(1+\lambda)N_z)-\theta \leq 2m(1+\lambda)N_z-z,$$

or equivalently, $2mN_{\lambda m}-\theta \leq -z$. The latter inequality holds as $\theta - z = 2m(1+\lambda)$ and $N_{\lambda m} \leq 1$. □

Recall that $z = y+x-x^* = y+x-\lambda m$. We next complete the proof of the Lemma 5 by showing that

$$\Delta(x,y) \leq e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2mN_x+2nN_y-\theta}{\theta}+\frac{1}{m}\right) \leq H(z).$$

The first inequality holds by (D.9). To prove the second inequality, we observe that

$$\begin{aligned}
&e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2mN_x+2nN_y-\theta}{\theta}+\frac{1}{m}\right)\\
&\leq e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2m\max\{N_{\lambda m+z}+(1+\lambda)N_0, N_{\lambda m}+(1+\lambda)N_z\}-\theta}{\theta}+\frac{1}{m}\right) && \text{(D.10)}\\
&\leq \max\left\{e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2m(N_{\lambda m+z}+1+\lambda)-\theta}{\theta}+\frac{1}{m}\right), e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2m(N_{\lambda m}+(1+\lambda)N_z)-\theta}{\theta}+\frac{1}{m}\right)\right\} && \text{(D.11)}\\
&\leq H(z),
\end{aligned}$$

where (D.10) holds due to the convexity property established by Claim 1, (D.11) holds by rearrangement of terms, and the last inequality holds because each of the expressions in the max are at most $H(d)$, by Claims 2 and 3. □

Lemma 6. *For every $x \leq x^*$ and $y \geq 0$, $\Delta(x,y) \leq H(y+x^*-x)$.*

Proof. Let $z = y+x^*-x$, $n = (1+\lambda)m$, and $\theta = m+n+x+y$. The proof considers two cases: either $x < x^*$ or $x = x^*$. First, we suppose that $x < x^*$. Then,

$$\begin{aligned}
\Delta(x,y) &= e^{\frac{z+1}{\sqrt{m}}}\left(\frac{m+x}{\theta}\right)+e^{\frac{z-1}{\sqrt{m}}}\left(\frac{y+n}{\theta}\right)-e^{\frac{z}{\sqrt{m}}} && \text{(D.12)}\\
&= e^{\frac{z}{\sqrt{m}}}\left(e^{\frac{1}{\sqrt{m}}}\frac{m+x}{\theta}+e^{-\frac{1}{\sqrt{m}}}\left(\frac{y+n}{\theta}\right)-1\right)\\
&\leq e^{\frac{z}{\sqrt{m}}}\left(\left(1+\frac{1}{\sqrt{m}}+\frac{1}{m}\right)\frac{m+x}{\theta}+\left(1-\frac{1}{\sqrt{m}}+\frac{1}{m}\right)\left(\frac{n+y}{\theta}\right)-1\right) && \text{(D.13)}\\
&= e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{m+x-n-y}{\theta}+\frac{1}{m}\right) = e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{-z}{\theta}+\frac{1}{m}\right) && \text{(D.14)}\\
&\leq e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{-z+2nN_z}{\theta}+\frac{1}{m}\right) = H(z). && \text{(D.15)}
\end{aligned}$$

To see why (D.12) holds, observe that $\frac{x+m\bar{N}_x}{\theta}$ and $\frac{N_x}{\theta}$ are the transition probabilities of $\widehat{\mathcal{N}}$ from the state (x,y) to the states $(x-1,y)$ and $(x,y+1)$, respectively; these transition probabilities sum up to $\frac{m+x}{\theta}$. When either of these transitions occur, the value of U changes from $e^{\frac{z}{\sqrt{m}}}$ to $e^{\frac{z+1}{\sqrt{m}}}$. Also, $\frac{nN_y}{\theta}$ and $\frac{y+n\bar{N}_y}{\theta}$ are the transition probabilities of $\widehat{\mathcal{N}}$ from the state (x,y) to the states $(x+1,y)$ and $(x,y-1)$, respectively; these transition probabilities sum up to $\frac{y+n}{\theta}$. (Note that the transition probabilities are 0 if $x-1$ or $y-1$ are negative.) When either of these transitions occur, the value of U changes from $e^{\frac{z}{\sqrt{m}}}$

to $e^{\frac{z-1}{\sqrt{m}}}$. Inequality (D.13) holds because $e^{\alpha} \leq 1+\alpha+\alpha^2$ holds for every real number in $[-1,1]$, and (D.15) holds because $N_z \geq 0$.

To complete the proof, it remains to prove the claim for the case of $x = x^*$. In this case, when a transition from (x,y) to $(x,y+1)$ occurs, the value of U changes from $e^{\frac{z}{\sqrt{m}}}$ to $e^{\frac{z+1}{\sqrt{m}}}$ (whereas in the above case, the value of U changes to $e^{\frac{z-1}{\sqrt{m}}}$ when this transition occurs). Accounting for this difference slightly changes the above calculations, but leads to the same conclusion:

$$\begin{aligned}
&\Delta(x,y)e^{\frac{z+1}{\sqrt{m}}}\left(\frac{m+x+nN_y}{\theta}\right)+e^{\frac{z-1}{\sqrt{m}}}\left(\frac{y+n\bar{N}_y}{\theta}\right)-e^{\frac{z}{\sqrt{m}}}\\
&\leq e^{\frac{z}{\sqrt{m}}}\left(\left(1+\frac{1}{\sqrt{m}}+\frac{1}{m}\right)\frac{m+x+nN_y}{\theta}+\left(1-\frac{1}{\sqrt{m}}+\frac{1}{m}\right)\left(\frac{y+n\bar{N}_y}{\theta}\right)-1\right)\\
&=e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{m+x+nN_y-y-n\bar{N}_y}{\theta}+\frac{1}{m}\right)\\
&=e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{-z+2nN_z}{\theta}+\frac{1}{m}\right)=H(z),
\end{aligned}$$

where the penultimate equality follows from $x = x^* = n-m$ and $y = z$. □

Lemma 7. *For every* $x,y \geq 0$ $\Delta(x,y) \leq H(y+|x-\lambda m|)$.

Proof. By Lemma 5, if $x > x^*$, then $\Delta(x,y) \leq H(y+x-x^*)$. By Lemma 6, if $x \leq x^*$, then $\Delta(x,y) \leq H(y+x^*-x)$. The claim follows immediately from the two latter bounds and the fact that $x^* = \lambda m$. □

Lemma 8. *For every* $k > 0$ *satisfying* $\sup_{z \geq k\sqrt{m}} H(z) < 0$ *it holds that*

$$\sum_{(x,y)\notin B_k} \hat{\rho}_{x,y} \leq -\frac{\sup\limits_{0\leq z\leq k\sqrt{m}} H(z)}{\sup\limits_{z\geq k\sqrt{m}} H(z)}.$$

Proof. Applying Proposition 9 on $\widehat{\mathcal{N}}$ directly implies that

$$\sum_{(x,y)\notin B_k} \hat{\rho}_{x,y} \leq \alpha+\frac{\beta}{\gamma} \tag{D.16}$$

holds if there exist $\alpha, \beta \geq 0$ and $\gamma > 0$ such that

$$\begin{aligned}
\forall (x,y) \notin B_k, \Delta(x,y) &\leq -\gamma f(x,y),\\
\forall (x,y) \in B_k, f(x,y) &\leq \alpha,\\
\forall (x,y) \in B_k, \Delta(x,y) &\leq \beta.
\end{aligned}$$

Since $f(x,y) = \mathbb{1}_{(x,y)\notin B_k}$ by definition, then we can set $\alpha = 0$. Recall that

$$B_k = \{(x,y) \in \mathbb{Z}_+^2 : |x-x^*|+|y| < k\sqrt{m}\}.$$

Hence, Lemma 7 implies that

$$\sup_{(x,y)\in B_k} \Delta(x,y) \leq \sup_{0\leq z\leq k\sqrt{m}} H(z).$$

Therefore, we can set

$$\beta = \sup_{0\leq z\leq k\sqrt{m}} H(z). \tag{D.17}$$

We note that $\beta > 0$ holds since $H(0) = e^{\frac{1}{\sqrt{m}}} - 1 > 0$.

Finally, we observe that by Lemma 7,

$$\sup_{(x,y)\notin B_k} \Delta(x,y) \leq \sup_{z\geq k\sqrt{m}} H(z).$$

Therefore, we can set

$$\gamma = -\sup_{z\geq k\sqrt{m}} H(z). \tag{D.18}$$

Since $\sup_{z\geq k\sqrt{m}} H(z) < 0$ holds by assumption, then $\gamma > 0$.

We have set $\alpha=0$, and have set β,γ by (D.17) and (D.18), respectively. This choice of parameters, together with (D.16), directly proves the claim. □

D.2.3. A Concentration Bound for ρ.

Corollary 1 (of Lemma 8) *For every $k>0$ satisfying $\sup_{z\geq k\sqrt{m}}\{H(z)\}<0$ it holds that*

$$\sum_{(x,y)\notin B_k}\rho_{x,y}\leq -\frac{\sup\limits_{0\leq z\leq k\sqrt{m}} H(z)}{\sup\limits_{z\geq k\sqrt{m}} H(z)}\cdot\frac{2(1+\lambda+k/\sqrt{m})}{2+\lambda}.$$

Proof. For a subset $S\in\mathbb{Z}_+^2$, let $\rho[S]$ denote $\sum_{(x,y)\in S}\rho_{x,y}$. Similarly, let $\hat{\rho}[S]$ denote $\sum_{(x,y)\in S}\hat{\rho}_{x,y}$. We denote $\mathbb{Z}_+^2\backslash S$ by $\overline{S}$.

If $\rho[\overline{B_k}]\leq\hat{\rho}[\overline{B_k}]$, then the claim holds by the upper bound on $\hat{\rho}[\overline{B_k}]$ provided by Lemma 8. So, suppose that this is not the case; i.e.,

$$\rho[B_k]<\hat{\rho}[B_k]. \tag{D.19}$$

Define $w_{x,y}=\frac{1}{m(2+\lambda)+x+y}$. We note that $w_{0,0}\geq w_{x,y}$ holds for every $(x,y)\in\mathbb{Z}_+^2$. Also, define $\underline{w}=w_{\lambda m+k\sqrt{m},k\sqrt{m}}$. We note that $\underline{w}\leq w_{x,y}$ for every $(x,y)\in B_k$.

By Fact 2, regarding the steady-state distribution of Embedded Markov chains, it holds that

$$\frac{\rho[\overline{B_k}]}{\rho[B_k]}\leq\frac{\hat{\rho}[\overline{B_k}]}{\hat{\rho}[B_k]}\cdot\frac{w_{0,0}}{\underline{w}}.$$

This implies that

$$\rho[\overline{B_k}]\leq\rho[B_k]\frac{\hat{\rho}[\overline{B_k}]}{\hat{\rho}[B_k]}\cdot\frac{w_{0,0}}{\underline{w}}.$$

The above inequality, together with (D.19), implies that

$$\rho[\overline{B_k}]\leq\hat{\rho}[\overline{B_k}]\cdot\frac{w_{0,0}}{\underline{w}}=\hat{\rho}[\overline{B_k}]\cdot\frac{2(1+\lambda+k/\sqrt{m})}{2+\lambda}.$$

The above bound, together with Lemma 8, implies that

$$\rho[\overline{B_k}]\leq -\frac{\sup\limits_{0\leq z\leq k\sqrt{m}} H(z)}{\sup\limits_{z\geq k\sqrt{m}} H(z)}\cdot\frac{2(1+\lambda+k/\sqrt{m})}{2+\lambda}.$$

□

Theorem 2. *For every $k>0$, define*

$$Q(k)=e^k\left(\frac{1}{\sqrt{m}}\frac{k\sqrt{m}-2m(1+\lambda)N_{k\sqrt{m}}}{2m(1+\lambda)+k\sqrt{m}}-\frac{1}{m}\right),$$

$$R(k)=\sup_{0\leq z\leq k\sqrt{m}} H(z),$$

$$\Phi(k)=\frac{R(k)}{Q(k)}\cdot\frac{2(1+\lambda+k/\sqrt{m})}{2+\lambda}.$$

Then, $\sum_{(x,y)\notin B_k}\rho_{x,y}\leq\Phi(k)$ holds if $Q(k)>0$.

Proof. First, we prove the following claim.

Claim 4. $\sup_{z\geq k\sqrt{m}}H(z)\leq -Q(k)$.

Proof. Observe that for every $z \ge k\sqrt{m}$,

$$
\begin{aligned}
H(z) &= -e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{z-2m(1+\lambda)N_z}{2m(1+\lambda)+z}-\frac{1}{m}\right)\\
&\le -e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\left(\frac{z}{2m(1+\lambda)+z}-\frac{2m(1+\lambda)N_z}{2m(1+\lambda)+z}\right)-\frac{1}{m}\right)\\
&\le -e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\left(\frac{k\sqrt{m}}{2m(1+\lambda)+k\sqrt{m}}-\frac{2m(1+\lambda)N_z}{2m(1+\lambda)+z}\right)-\frac{1}{m}\right) \qquad \text{(D.20)}\\
&\le -e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\left(\frac{k\sqrt{m}}{2m(1+\lambda)+k\sqrt{m}}-\frac{2m(1+\lambda)N_{k\sqrt{m}}}{2m(1+\lambda)+k\sqrt{m}}\right)-\frac{1}{m}\right) \qquad \text{(D.21)}\\
&\le -e^{k}\left(\frac{1}{\sqrt{m}}\left(\frac{k\sqrt{m}}{2m(1+\lambda)+k\sqrt{m}}-\frac{2m(1+\lambda)N_{k\sqrt{m}}}{2m(1+\lambda)+k\sqrt{m}}\right)-\frac{1}{m}\right) \qquad \text{(D.22)}\\
&= -Q(k).
\end{aligned}
$$

where (D.20) holds by Fact 5, (D.21) holds by the fact that $z \ge k\sqrt{m}$, and (D.22) holds since $Q(k) > 0$. □

By the above claim, $\sup_{z \ge k\sqrt{m}} H(z) < 0$. Thus, Corollary 1 applies, which implies that

$$
\sum_{(x,y)\notin B_k}\rho_{x,y} \le -\frac{\sup\limits_{0\le z\le k\sqrt{m}} H(z)}{\sup\limits_{z\ge k\sqrt{m}} H(z)}\cdot\frac{2(1+\lambda+k/\sqrt{m})}{2+\lambda} \le \frac{R(k)}{Q(k)}\cdot\frac{2(1+\lambda+k/\sqrt{m})}{2+\lambda},
$$

where the last inequality holds by Claim 4. This completes the proof. □

D.2.4. A concentration bound for π.

Lemma 9. *For every $\alpha \ge 0$ and $z \le 2\alpha(1+\lambda)\sqrt{m}$, it holds that $H(z) \le e^{2\alpha(1+\lambda)}\left(\frac{1}{\sqrt{m}}+\frac{1}{m}\right)$.*

Proof. Observe that

$$
H(z) = e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2m(1+\lambda)N_z-z}{2m(1+\lambda)+z}+\frac{1}{m}\right) \le e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}+\frac{1}{m}\right) \le e^{2\alpha(1+\lambda)}\left(\frac{1}{\sqrt{m}}+\frac{1}{m}\right), \qquad \text{(D.23)}
$$

where the last inequality holds because $z \le 2\alpha(1+\lambda)\sqrt{m}$. □

Lemma 10. *For every $\alpha \ge 3$, $m \ge \max\{36, p^{-2}\}$, and $z \ge 2\alpha(1+\lambda)\sqrt{m}$, $H(z) \le \frac{1}{\sqrt{m}} - \frac{1}{m}e^{\frac{z}{\sqrt{m}}}$.*

Proof. Observe that

$$
\begin{aligned}
H(z) &= e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2m(1+\lambda)N_z-z}{2m(1+\lambda)+z}+\frac{1}{m}\right) \le e^{\frac{z}{\sqrt{m}}}\left(\frac{1}{\sqrt{m}}\frac{2m(1+\lambda)e^{-pz}-z}{2m(1+\lambda)+z}+\frac{1}{m}\right) \qquad \text{(D.24)}\\
&\le \frac{1}{\sqrt{m}}\frac{2m(1+\lambda)e^{\frac{z}{\sqrt{m}}-pz}}{2m(1+\lambda)+z}-\frac{1}{\sqrt{m}}\frac{ze^{\frac{z}{\sqrt{m}}}}{2m(1+\lambda)+z}+\frac{1}{m}e^{\frac{z}{\sqrt{m}}}\\
&\le \frac{1}{\sqrt{m}}+e^{\frac{z}{\sqrt{m}}}\left(-\frac{1}{\sqrt{m}}\frac{z}{2m(1+\lambda)+z}+\frac{1}{m}\right) \qquad \text{(D.25)}\\
&\le \frac{1}{\sqrt{m}}+e^{\frac{z}{\sqrt{m}}}\left(-\frac{1}{\sqrt{m}}\frac{2\alpha(1+\lambda)\sqrt{m}}{2m(1+\lambda)+2\alpha(1+\lambda)\sqrt{m}}+\frac{1}{m}\right) \qquad \text{(D.26)}\\
&\le \frac{1}{\sqrt{m}}+e^{\frac{z}{\sqrt{m}}}\left(-\frac{1}{m/\alpha+\sqrt{m}}+\frac{1}{m}\right) \le \frac{1}{\sqrt{m}}-\frac{1}{m}e^{\frac{z}{\sqrt{m}}} \qquad \text{(D.27)}
\end{aligned}
$$

where (D.24) holds because $1-\alpha \le e^{-\alpha}$ for every real α, (D.25) holds because $\frac{z}{\sqrt{m}} - pz \le 0$ (which holds since $m \ge p^{-2}$), (D.26) holds by Fact 5, and (D.27) holds because $\alpha \ge 3$ and $m \ge 36$. □

Corollary 2. *For every $m \ge \max\{36, p^{-2}\}$ and every $z \ge 0$, it holds that $H(z) \le e^{6(1+\lambda)}\left(\frac{1}{\sqrt{m}}+\frac{1}{m}\right)$.*

Proof. Let $\alpha=3$. Lemma 9 proves the claim for when $z\le 2\alpha(1+\lambda)\sqrt{m}$. On the other hand, when $z\ge 2\alpha(1+\lambda)\sqrt{m}$, then Lemma 10 implies that

$$H(z)\le \frac{1}{\sqrt{m}}-\frac{1}{m}e^{\frac{z}{\sqrt{m}}}<\frac{1}{\sqrt{m}}.$$

□

Theorem 3 (Large market lower concentration bound) *There exist positive constants m_0,k_0,c_0 such that for every $m>m_0$ and $k>k_0$,*

$$\sum_{x=0}^{\lambda m-k\sqrt{m}}\pi_x\le c_0mke^{-k}.$$

Proof. Let $m_0=\max\{36,p^{-2}\}$ and $k_0=\max\{6(1+\lambda),\log m\}$. By Lemma 10, for every $m\ge m_0$ and $z\ge k_0\sqrt{m}$, it holds that $H(z)\le\frac{1}{\sqrt{m}}-\frac{1}{m}e^{\frac{z}{\sqrt{m}}}$. Hence, for every $k>k_0$ and $m>m_0$,

$$H(k\sqrt{m})\le\frac{1}{\sqrt{m}}-\frac{1}{m}e^{k}\le\frac{1}{\sqrt{m}}-1<0.$$

Consequently, Corollary 1 implies that for every $m>m_0$ and $k>k_0$,

$$\sum_{(x,y)\notin B_k}\rho_{x,y}\le-\frac{\sup\limits_{0\le z\le k\sqrt{m}}H(z)}{\sup\limits_{z\ge k\sqrt{m}}H(z)}\cdot\frac{2(1+\lambda+k/\sqrt{m})}{2+\lambda}. \tag{D.28}$$

By Corollary 2, for every $m>m_0$ and $z\ge 0$, $H(z)\le e^{6(1+\lambda)}$ holds, which implies that

$$\sup_{0\le z\le k\sqrt{m}}H(z)\le e^{6(1+\lambda)}. \tag{D.29}$$

On the other hand, Lemma 10 implies that for every $m>m_0$, $k>k_0$, and $z\ge k\sqrt{m}$, it holds that $H(z)\le\frac{1}{\sqrt{m}}-\frac{1}{m}e^{\frac{z}{\sqrt{m}}}$, which implies that

$$\sup_{z\ge k\sqrt{m}}H(z)\le\frac{1}{\sqrt{m}}-\frac{1}{m}e^{k}\le-\frac{1}{2m}e^{k}, \tag{D.30}$$

where the last inequality holds because $\frac{1}{\sqrt{m}}\le\frac{1}{2m}e^{k}$ holds when $m>m_0$. We next observe that (D.28), (D.29), and (D.30) together imply that

$$\sum_{(x,y)\notin B_k}\rho_{x,y}\le\frac{e^{6(1+\lambda)}}{\frac{1}{2m}e^{k}}\cdot\frac{2(1+\lambda+k/\sqrt{m})}{2+\lambda}\le 2e^{6(1+\lambda)}me^{-k}4k=8e^{6(1+\lambda)}kme^{-k},$$

where the last inequality holds because $\frac{2(1+\lambda+k/\sqrt{m})}{2+\lambda}\le 4k$ for every $k>k_0$.

The above bound, together with Lemma 4, implies that

$$\sum_{x=0}^{\lambda m-k\sqrt{m}}\pi_x\le\sum_{x=0}^{\lambda m-k\sqrt{m}}\rho_x\le\sum_{(x,y)\notin B_k}\rho_{x,y}\le 8e^{6(1+\lambda)}kme^{-k}.$$

Setting $c_0=8e^{6(1+\lambda)}$ concludes the proof. □

We next provide an upper concentration bound for the greedy policy.

D.3. *An upper concentration bound for the greedy policy*

In this section, we complement Theorem 3 by providing an upper concentration bound for the greedy process. The main idea is defining a simpler process that is coupled with $\mathcal{M}$, namely $\mathcal{M}_{\mathrm{u}}$. This process is defined such that the number of unmatched H agents in $\mathcal{M}$ is stochastically dominated by the number of unmatched H agents in $\mathcal{M}_{\mathrm{u}}$.

D.3.1. Definition of $\mathcal{M}_{\mathrm{u}}$. Consider a Markov process which is the same as $\mathcal{M}$ but with the following differences:

1. Matches are made only between H agents and E agents.
2. Matches are made greedily only upon the arrival of E agents.

3. E agents do not stay in the market: If upon the arrival of an E agent she is not matched to an H agent, then the E agent departs the market immediately.

The number of H agents in this random process is a one-dimensional Markov chain $\mathcal{M}_{\mathrm{u}}$. The state space of $\mathcal{M}_{\mathrm{u}}$ is $V(\mathcal{M}_{\mathrm{u}})=\{0,1,2,\ldots\}$. The Markov chain is at state x when the number of H agents present in the pool is x. The transition rate from a state x to state $x+1$ is $r_x=m(1+\lambda)$, since H agents arrive at a rate of $m(1+\lambda)$. The transition rate from a state x to state $x-1$ (when it exists) is $l_x=m(1-N_x)+x$. The first summand corresponds to the event that an arriving E agent is compatible to an H agent, and the second corresponds to the departures of H agents.

Since $\mathcal{M}_{\mathrm{u}}$ is irreducible and positive recurrent, it has a unique stationary distribution (Norris (1997), Theorem 3.5.3), which we denote by π^{u}. Let π_i^{u} be the probability that π^{u} assigns to state i.

Lemma 11. *The steady-state distribution of the number of H agents in $\mathcal{M}_{\mathrm{u}}$ stochastically dominates the steady-state distribution of the number of H agents in $\mathcal{M}$.*

The proof of the lemma is technical and is deferred to the Supplementary Appendix, Section i. The proof idea is defining a *coupling* of $\mathcal{M}$ and $\mathcal{M}_{\mathrm{u}}$ such that, in the coupled process, there are fewer H agents in the pool under $\mathcal{M}$ than under $\mathcal{M}_{\mathrm{u}}$ at any time, and more E agents.

Lemma 12. *Suppose that $m\geq 4(1+\lambda)^2$ and $p\lambda\geq\frac{\ln m}{2m}$. Then, for every positive integer k we have*

$$\pi^{\mathrm{u}}_{\lambda m+k}\leq e^{-\frac{k}{\sqrt{m}}+(3+2\lambda)}.$$

Proof. For notational simplicity, we denote π^{u} by η throughout this proof. We start by writing the balance equations, according to which, for every positive integer i, $\frac{\eta_i}{\eta_{i-1}}=\frac{r_i}{l_{i-1}}$.

Let $x^*=\lambda m$, and suppose $i\geq x^*$. Then,

$$\frac{\eta_i}{\eta_{i-1}}=\frac{(1+\lambda)m}{i+m(1-N_i)}=\frac{1+\lambda}{i/m+1-N_i}.$$

When $i\geq x^*+2(1+\lambda)\sqrt{m}+me^{-p\lambda m}$, and $m\geq 4(1+\lambda)^2$, we can write

$$\frac{\eta_i}{\eta_{i-1}}=\frac{1+\lambda}{i/m+1-N_i}\leq\frac{1+\lambda}{1+\lambda+(i-x^*)/m-e^{-p\lambda m}}\leq\frac{1+\lambda}{1+\lambda+2(1+\lambda)/\sqrt{m}} \tag{D.31}$$

$$\leq\frac{1}{1+2/\sqrt{m}}\leq 1-\frac{1}{\sqrt{m}}, \tag{D.32}$$

where (D.31) uses the fact that $N_i\leq e^{-ip}$ and that $i\geq 2(1+\lambda)\sqrt{m}+me^{-p\lambda m}$, and (D.32) uses the fact that $\frac{1}{1+\alpha}\leq 1-\alpha/2$ for every positive $\alpha\leq 1$.

For every positive integer k, we have $\frac{\eta_{x^*+k}}{\eta_{x^*}}=\prod_{j=0}^{k-1}\frac{\eta_{x^*+j+1}}{\eta_{x^*+j}}$. Then, for every integer $k\geq 2(1+\lambda)\sqrt{m}+me^{-p\lambda m}$, we can use (D.32) to write

$$\eta_{x^*+k}\leq\frac{\eta_{x^*+k}}{\eta_{x^*}}\leq e^{-\frac{k-2(1+\lambda)\sqrt{m}-me^{-p\lambda m}}{\sqrt{m}}}\leq e^{-\frac{k}{\sqrt{m}}+(3+2\lambda)}, \tag{D.33}$$

where the last inequality holds because $p\lambda m\geq\frac{\ln m}{2}$.

On the other hand, for every non-negative integer $k<2(1+\lambda)\sqrt{m}+me^{-p\lambda m}$, we have

$$\eta_{x^*+k}\leq 1<e^{-\frac{k-2(1+\lambda)\sqrt{m}-me^{-p\lambda m}}{\sqrt{m}}}\leq e^{-\frac{k}{\sqrt{m}}+(3+2\lambda)}, \tag{D.34}$$

where the last inequality holds because $p\lambda m\geq\frac{\ln m}{2}$. Finally, (D.33) and (D.34) conclude the proof. □

Theorem 4 (Large market upper concentration bound) *Suppose that $m\geq 4(1+\lambda)^2$ and $p\lambda\geq\frac{\ln m}{2m}$. Also, let π denote the steady-state distribution of $\mathcal{M}$. Then, for every positive integer k we have*

$$\sum_{j=k}^{\infty}\pi_{\lambda m+j}\leq\frac{m}{\sqrt{m}-1}e^{-\frac{k}{\sqrt{m}}+3+2\lambda}.$$

Proof. By Lemma 12, for every positive integer j we have $\pi^{\mathrm{u}}_{\lambda m+j}\leq e^{-\frac{j}{\sqrt{m}}+(3+2\lambda)}$. On the other hand, by Lemma 11, the steady-state distribution of the number of H agents in $\mathcal{M}_{\mathrm{u}}$ stochastically dominates the steady-state distribution of the number of H agents in $\mathcal{M}$. Therefore,

$$\sum_{j=k}^{\infty}\pi_{\lambda m+j}\leq\sum_{j=k}^{\infty}\pi^{\mathrm{u}}_{\lambda m+j}\leq\sum_{j=k}^{\infty}e^{-\frac{j}{\sqrt{m}}+(3+2\lambda)}=e^{3+2\lambda}\,\frac{e^{-\frac{k}{\sqrt{m}}}}{1-e^{-\frac{1}{\sqrt{m}}}}\leq e^{3+2\lambda}\,\frac{e^{-\frac{k}{\sqrt{m}}}}{\frac{1}{\sqrt{m}}-\frac{1}{m}},$$

where the last inequality holds because $e^z\leq 1+z+z^2$ for every real number in $[-1,1]$. □

E. MATCH RATE AND WAITING TIME UNDER THE GREEDY POLICY

E.1. *Match rates under the greedy policy*

Claim 5. $q_E^G(m)\geq 1-O(1/m)$.

Proof. Fix an E agent, namely a, who has just arrived to the market. Let x_a, y_a respectively denote the size of the H pool and the E pool just before a arrives. By the PASTA property of the Poisson process,[38] the probability distribution of (x_a,y_a) is the same as the steady-state distribution π. Therefore, theorem 3 implies that for sufficiently large m,

$$\mathbb{P}_\pi\left[x_a<m-3\log m\sqrt{m}\right]\leq e^{-3\log m}3c_0 m\log m\leq m^{-1}.$$

This implies that, upon her arrival, agent a has a compatible H agent with probability at least

$$(1-m^{-1})\left(1-(1-p)^{\lambda m-3\log m\sqrt{m}}\right)=1-O(m^{-1}). \tag{E.35}$$

This probability is a lower bound for $q_E^G(m)$. □

Claim 6. $q_H^G(m)\in\left(\frac{1}{1+\lambda}-O(m^{-1/3}),\frac{1}{1+\lambda}+O(m^{-1/3})\right)$.

Proof. Fix $m>0$. Suppose that the market starts at time 0 when there are no agents in the market. For any $t>0$, let $m_E(t), m_H(t)$, respectively denote the number of E agents and H agents that arrive from time 0 to time t. Also, let $\psi_H(t)$ denote the number of H agents that are matched from time 0 to time t.

By the Ergodic theorem, $\lim_{t\to\infty}\psi_H(t)/m_H(t)=q_H^G(m)$. Hence, it suffices to prove the claim for the left-hand side of the equality. Since E agents arrive according to a Poisson process with rate m, then, for any $t>1$, the event $m_E(t)\in[mt-(mt)^{2/3},mt+(mt)^{2/3}]$ holds with very high probability. This holds by the concentration bound of Fact 3 for the Poisson distribution. Similarly, since H agents arrive according to a Poisson process with rate $(1+\lambda)m$, for any $t>1$, the event

$$m_H(t)\in[(1+\lambda)mt-((1+\lambda)mt)^{2/3},(1+\lambda)mt+m^{2/3}]$$

holds with very high probability due to Fact 3. This implies that, for any $t>1$,

$$\frac{\psi_H(t)}{m_H(t)}\leq\frac{m_E(t)}{m_H(t)}\leq\frac{mt+(mt)^{2/3}}{(1+\lambda)mt-((1+\lambda)mt)^{2/3}} \tag{E.36}$$

holds with very high probability.

Next, we provide a lower bound for $\frac{\psi_H(t)}{m_H(t)}$. Recall from (E.35) that, upon her arrival, any E agent is matched to an H agent with probability at least $1-O(m^{-1})$. Therefore,

$$\frac{\psi_H(t)}{m_H(t)}\geq\frac{m_E(t)(1-O(m^{-1}))}{(1+\lambda)mt+((1+\lambda)mt)^{2/3}}\geq\frac{(mt-(mt)^{2/3})(1-O(m^{-1}))}{(1+\lambda)mt+((1+\lambda)mt)^{2/3}}. \tag{E.37}$$

Now observe that, for any $t>1$, (E.36) and (E.37) together imply that

$$\frac{\psi_H(t)}{m_H(t)}\in\left(\frac{1}{1+\lambda}-O(m^{-1/3}),\frac{1}{1+\lambda}+O(m^{-1/3})\right)$$

holds with very high probability. Since, by the Ergodic theorem, $\lim_{t\to\infty}\psi_H(t)/m_H(t)$ exists and converges to $q_H^G(m)$ almost surely in any sample path, then we have

$$q_H^G(m)\in\left(\frac{1}{1+\lambda}-O(m^{-1/3}),\frac{1}{1+\lambda}+O(m^{-1/3})\right).$$

□

[38] PASTA, or Poisson Arrivals See Time Averages, is a well-known property in the queuing literature; e.g., see Harchol-Balter (2013).

Lemma 13. *Under the greedy policy, the match rate of hard-to-match agents $q_H^G(m)$ is $\frac{1}{1+\lambda}-o(1)$ and the match rate of easy-to-match agents $q_E^G(m)$ is $1-o(1)$.*

Proof. The lemma follows immediately from Claims 5 and 6. □

E.2. *Distribution of waiting time under greedy matching*

We show that as m approaches infinity, the waiting time for easy-to-match agents converges in distribution to the degenerate distribution at 0, and the waiting time for hard-to-match agents converges to the exponential distribution with rate $1/d+1/\lambda$.

Lemma 14. *Under the greedy policy, as m approaches infinity, the waiting time of an easy-to-match agent converges in distribution to the degenerate distribution at* 0.

Proof. Fix an E agent, e, and let w_e denote the waiting time for e. For any fixed constant $t>0$, we will show that $\lim_{m\to\infty}\mathbb{P}[t>w_e]=1$. This will prove the claim. Recall from (E.35) that upon her arrival, agent e is matched to an H agent with probability $1-O(m^{-1})$. Therefore, $\mathbb{P}[w_e=0]=1-O(m^{-1})$, which implies that, $\lim_{m\to\infty}\mathbb{P}[t>w_e]=1$ holds for any $t>0$. □

Lemma 15. *As m approaches infinity, the waiting time of hard-to-match agents converges in distribution to the exponential distribution with rate $\frac{1}{d}+\frac{1}{\lambda}$.*

We sketch the proof below. The formal proof is technical and is presented in the Supplementary Appendix, Section i. We will use $\mathsf{Exp}(x)$ to denote the exponential distribution with rate x.

Proof sketch We define a new process, namely $\mathcal{P}$, in which there are no easy-to-match agents. Instead, *attach* an exponential clock to each hard-to-match agent which ticks at rate $1/\lambda$. We call this clock the *match clock* of the agent. We consider an agent to be matched if the match clock ticks before the agent becomes critical. Without providing a formal proof in this proof sketch, we suppose that H agents in the new process $\mathcal{P}$ have approximately the same waiting time as in the original process (the greedy policy). Given this assumption, we compute the distribution for the waiting time of a hard-to-match agent h in $\mathcal{P}$.

Consider the agent h and suppose she has entered the pool at time t_0. Note that h is matched if and only if it is matched before her criticality clock ticks. Let t_1,t_2 be random variables such that $t_1\sim\mathsf{Exp}(1/\lambda), t_2\sim\mathsf{Exp}(1/d)$. These random variables are interpreted as follows. The agent becomes critical at time t_0+t_2 if she is not matched by then, i.e., if the match clock attached to her has not ticked by then. The agent's match clock ticks at time t_0+t_1. So, the agent is matched if and only if $t_1<t_2$. Alternatively, we can say the agent is matched if and only if $t_1=t_{\min}$ where $t_{\min}=\min\{t_1,t_2\}$. Since $t_{\min}$ is distributed according to $\mathsf{Exp}(1/d+1/\lambda)$, and since $t_{\min}$ equals the waiting time of the agent, the claim is proved. □

E.3. *Proof of Proposition 5*

Proof of Proposition 5 The claim about the match rate was proved in Lemma 13, where we showed that under the greedy policy, the match rate of hard-to-match agents $q_H^G(m)$ is $\frac{1}{1+\lambda}-O(\frac{1}{(1+\lambda)\sqrt{m}})$ and the match rate of easy-to-match agents $q_E^G(m)$ is $1-o(1)$. The claim about waiting times was proved in Lemmas 14 and 15, for easy- and hard-to-match agents, respectively. □

F. ANALYSIS OF THE BATCHING POLICY

F.1. *Preliminary graph theory results*

For every graph G, we let $V(G)$ denote the set of its nodes and $E(G)$ denote the set of its edges. An *independent* set in a graph G is a subset of nodes $S\subseteq V(G)$ such that no two nodes in S are adjacent (i.e. are connected by an edge) in G.

We denote a bipartite graph by $G(X,Y)$ where X,Y denote the set of nodes on each side of G. (That is, $V(G)=X\cup Y$, and both X,Y are independents sets in G.) A *matching* is a set of edges such that no two of the edges have a common node. The size of a matching is the number of the edges that it contains. A *perfect matching* in $G(X,Y)$ is a matching with size $\min\{|X|,|Y|\}$.

Lemma 16. *Let $G(X,Y)$ be a randomly drawn bipartite graph with non-random X,Y being its partitions where $|X|=|Y|=n$. Suppose that the probability that a node $u\in X$ is connected a node $v\in Y$ equals $p\in(0,1)$ independently across all pairs (u,v). Then, the graph contains a perfect matching with probability at least $1-n2^{2n}p^{n^2/4}$.*

Proof. By the König–Egerváry Theorem, there exists a perfect matching in G if and only if the size of the maximum independent set is at most n (West, 2000).

For $X'\subseteq X$ and $Y'\subseteq Y$, let $E(X',Y')$ denote the event in which no node in X' is adjacent to a node in Y'. We note that if $E(X',Y')$ happens then $X'\cup Y'$ is an independent set. Also, let the set $\mathcal{E}'$ be the set of all pairs (X',Y') such that $X'\subseteq X, Y'\subseteq Y$, and $|X'|+|Y'|=n+1$. Therefore, $\bigcup_{(X',Y')\in\mathcal{E}'} E(X',Y')$ is the event that there exists an independent set of size larger than n (which also means that no perfect matching exists). By a union bound, the probability that a perfect matching does not exist in G is then at most

$$\mathbb{P}\left[\bigcup_{(X',Y')\in\mathcal{E}'} E(X',Y')\right]\leq \sum_{(X',Y')\in\mathcal{E}'} \mathbb{P}\big[E(X',Y')\big]. \tag{F.38}$$

Consider (X',Y') with $|X'|+|Y'|=n+1$, and let $i=|X'|$. The probability that $E(X',Y')$ holds then equals $p^{i(n-i+1)}$. Therefore,

$$\sum_{(X',Y')\in\mathcal{E}'} \mathbb{P}\big[E(X',Y')\big]=\sum_{i=1}^{n}\binom{n}{i}\binom{n}{n-i+1}(1-p)^{i(n-i+1)}=\sum_{i=1}^{n}\binom{n}{i}\binom{n}{i-1}(1-p)^{i(n-i+1)}$$

$$\leq\sum_{i=1}^{n}2^{2n}(1-p)^{i(n-i+1)}\leq n2^{2n}(1-p)^{n^2/4},$$

where the penultimate inequality follows from the fact that the product of the binomial coefficients $\binom{n}{i}\binom{n}{i-1}$ is bounded by 2^{2n}, as each of the multiplicands is bounded by 2^n. □

Corollary 3 (Corollary of Lemma 16) *Let $G(X,Y)$ be a random bipartite graph with X,Y being its partitions where $|Y|=n$ and $|X|<|Y|$. A node $u\in X$ is adjacent to a node $v\in Y$ with probability p, independently across all pairs (u,v). Then, G contains a matching of size $|X|$ with probability at least $1-n2^{2n}p^{n^2/4}$.*

Proof. Construct a graph H from G by adding $n-|X|$ dummy nodes to X. Let each dummy node x' and each node $y\in Y$ be adjacent independently with probability p.

Let $\mathfrak{p}$ denote the probability that H contains a matching of size n, and $\mathfrak{q}$ denote the probability that G contains a matching of size $|X|$. As any matching of size $n=|Y|$ in H must cover every node in X, then $\mathfrak{q}>\mathfrak{p}$. By Lemma 16, $\mathfrak{p}\geq 1-n2^{2n}p^{n^2/4}$. This concludes the proof. □

Definition 13. *In a bipartite graph $G(X,Y)$, a subset $S\subseteq X\cup Y$ is called an (x,y)-independent set if S is an independent set in G such that $|S\cap X|=x$ and $|S\cap Y|=y$.*

Lemma 17. *Let $\alpha,\beta,\gamma>0$ be arbitrary constants such that $\alpha,\beta\in(0,1)$. Let $G(X,Y)$ be a bipartite graph such that $|Y|=\gamma|X|$ and, furthermore, for every pair of nodes $u\in X$ and $v\in Y$, u is adjacent to v independently with probability $p>0$. Then, with high probability as $|X|$ grows large, G contains no $(\alpha|X|,\beta|Y|)$-independent set.*

Proof. Consider arbitrary subsets $X'\subseteq X$ and $Y'\subseteq Y$ such that $|X'|=\alpha|X|$ and $|Y'|=\beta|Y|$. The probability that $X'\cup Y'$ is an independent set is $(1-p)^{\alpha\beta|X|\cdot|Y|}$. Hence, the probability that there exists at least one $(|X'|,|Y'|)$-independent set in G is bounded by

$$\binom{|X|}{|X'|}\binom{|Y|}{|Y'|}(1-p)^{\alpha\beta|X|\cdot|Y|}\leq 2^{|X|(1+\gamma)}(1-p)^{\alpha\beta\gamma|X|^2}.$$

The right-hand side of the above inequality approaches 0 as $|X|$ approaches infinity. □

F.2. *Preliminary definitions and lemmas*

In the analysis, we suppose that time is indexed by non-negative real numbers. The batching policy makes matches at times iT for every positive integer i; these times are called *matching times*. For every $i\geq 0$, the interval $(iT,(i+1)T]$ is called *period* i.

The batching policy *executes* a matching at the *end* of every period i: at time $(i+1)T$, it finds the largest matching in the pool. If there are several such matchings, it selects the matching among them which has the maximum number of H agents. The policy then executes the selected matching and removes the agents involved in that matching from the pool.

For any matching time t, x_t and y_t, respectively denote the number of H agents and the number of E agents in the pool after the execution of the matching at time t. If t is not a matching time, let x_t and y_t denote the number of H and E agents in the pool at time t, respectively.

We note that the sequence $\langle(x_{iT}, y_{iT})\rangle_{i\geq 0}$ is a discrete-time Markov chain with a state space $\mathbb{Z}_+^2$. Since this Markov chain is Ergodic, it has a steady-state distribution. Since we are performing a steady-state analysis, we suppose that (x_0, y_0) is drawn from the steady-state distribution. This assumption is without loss of generality by the Ergodic theorem for Markov chains.

Lemma 18. *The match rate of agents of type* Θ *equals* $1-w_{\Theta}^{\tau}(m)/d$ *if* τ *is either the batching or greedy policy.*

The proof of Lemma 18 is identical to the proof of Lemma 1 adjusting for the fact that instead of an upper bound on the match rate we know its exact value; i.e., the probability that the agent is matched is given by $1-\frac{1}{d}\mathbb{E}_{\alpha_i}[\varphi_i]=\mathbb{E}_{\alpha_i}[\mu_i]$.

Lemma 19. *Consider a time interval* (a,b) *and let* $c=b-a$. *Conditional on an agent arriving in the interval* (a,b), *the criticality time of the agent is larger than b with probability* $\gamma_{c,d}=\frac{1-e^{-c/d}}{c/d}$.

Proof. By the properties of the Poisson process, the distribution of the arrival time of an agent conditional on the agent arriving in an interval $[a,b]$ equals the uniform distribution over the interval $[a,b]$. Hence, the chance that the criticality time of the agent is larger than b equals

$$\int_0^c \frac{1}{c}e^{-(c-s)/d}\mathrm{d}s=\frac{d(1-e^{-c/d})}{c}. \qquad \square$$

F.3. *Analysis of match rate*

We first provide an upper bound for match rate, and then a matching lower bound for it.

Lemma 20. *Under a batching policy with batch length* T, $q_E^B(m)\leq\gamma_{T,d}$ *and* $q_H^B(m)\leq\frac{\gamma_{T,d}}{1+\lambda}$.

Proof. Let $i>0$ be an arbitrary integer. Conditional on an E agent arriving at a time in the interval $(iT,(i+1)T]$, the agent is present in the pool at time $(i+1)T$ with probability $\gamma_{T,d}$ by Lemma 19. Therefore, the match rate of E agents is at most $\gamma_{T,d}$.

To prove the claim for H agents, observe that H agents can be matched only to E agents. Under the batching policy, only a fraction $\gamma_{T,d}$ of the E agents would not become critical before the first matching time after their arrival. Hence, the match rate of H agents is at most $\frac{\gamma_{T,d}}{1+\lambda}$. □

To provide lower bounds on match rate, we need the following definitions and lemmas.

Definition 14. *For an integer* $i\geq 0$, *an agent present in the pool at time* $(i+1)T$ *is called a* new *agent if she has arrived later than time* iT.

Definition 15. *In a graph G whose nodes correspond to E and H agents, an edge* (u,v) *is a* cross-edge *if* u,v *are agents of different types.*

Lemma 21. *Let* e_i *denote the number of new E agents who are present in the pool at time* $(i+1)T$. *Then, the matching executed at the matching time* $(i+1)T$ *involves at least* e_i *cross-edges, whp.*

Proof. We first construct a bipartite graph $G(X,Y)$, where X and Y, respectively denote the set of H agents in the pool at time $(i+1)T$ before the matching is executed, and the set of new E agents in the pool at time $(i+1)T$ before the matching is executed.

Claim 7. *Whp, it holds that* $|Y|<(1+\lambda/2)\gamma_{T,d}mT<|X|$.

Proof. By Lemma 19, conditional on an agent arriving to the pool after time iT, that agent remains in the pool until time $(i+1)T$ with probability $\gamma_{T,d}$, independently. (The independence is due to the independence of the criticality times.)

Therefore, the random variable $|Y|$ is a Poisson random variables with mean $\gamma_{T,d}mT$. This holds because E agents arrive with rate m but are present in the pool in the next batching time after their arrival only with probability $\gamma_{T,d}$. This fact, together with the concentration bound of Fact 3 for the Poisson distribution, implies that $|Y| < (1+\lambda/2)\gamma_{T,d}mT$ holds whp.

Let X' denote the set of new H agents in the pool at time $(i+1)T$ before the matching is executed. By Lemma 19, $|X'|$ is a Poisson random variables with mean $\gamma_{T,d}(1+\lambda)mT$. (This holds by the same argument for the case of E agents, with the difference that the arrival rate of H agents is $(1+\lambda)m$.) This fact, together with the concentration bound of Fact 3 for the Poisson distribution, implies that $|X'| > (1+\lambda/2)\gamma_{T,d}mT$ holds whp. Since $X' \subseteq X$, therefore, $|X| > (1+\lambda/2)\gamma_{T,d}mT$ holds whp. The proof is complete. □

Recall that $e_i = |Y|$. By Claims 7 and 3, whp there exists a matching of size $|Y|$ in G. Given this fact, the next claim concludes the proof.

Claim 8. *If there exists a matching of size $|Y|$ in G, then the matching executed at time $(i+1)T$ involves at least $|Y|$ cross-edges.*

Proof. Let M denote a matching of size $|Y|$ in G. Let M' denote the maximum matching chosen by the batching policy to be executed at time $(i+1)T$. We construct a graph, G', where $V(G')$ is the set of all of the agents present in the pool at time $(i+1)T$, before the matching is executed, and $E(G')=E(M)\cup E(M')$. Thus, G' must be a union of paths and cycles (West, 2000). Let C, P_e, P_o, respectively denote the set of cycles, the set of paths of even length, and the set of paths of odd length in G'.

For a subgraph F of G', let $D(F)$ denote the set of cross-edges of F. For two subgraphs F_1, F_2 of G', let $F_1 \Delta F_2$ denote the subgraph of G' with the set of edges $E(F_1)\cup E(F_2)-(E(F_1)\cap E(F_2))$.

First, we show that for every cycle or path of even length $Z \in C \cup P_e$,

$$|D(Z)\cap E(M)| \leq |D(Z)\cap E(M')|. \tag{F.39}$$

Suppose not. Then, observe that $M'\Delta Z$ would be a matching with the same size as M' but a larger number of cross-edges. This contradicts the definition of M'. Hence, (F.39) must hold.

Next, we show that for every path of odd length $Z \in P_o$,

$$|D(Z)\cap E(M)| \leq |D(Z)\cap E(M')|. \tag{F.40}$$

To see why, first note that the first and last edges in Z must belong to M'. Otherwise $M'\Delta Z$ would be a matching with a larger size than M', which would be a contradiction. Now, consider a cross-edge (e,h) belonging to both Z and M, where e,h are respectively E and H agents. Since the first and last edges in Z must belong to M', then there must exist a cross-edge (e',h) belonging to M'. This means that for every cross-edge (e,h) belonging to both Z and M, there exists a cross-edge (e',h) belonging to both Z and M'. Therefore, (F.40) holds.

Finally, (F.39) and (F.40) together imply that the number of cross-edges in M' is at least as large as the number of cross-edges in M, which equals $|Y|$. □

This completes the proof of Lemma 21. □

Lemma 22. *For a batching policy with batch length T, $q_E^B(m) \geq \gamma_{T,d} - o(1)$ and $q_H^B(m) \geq \frac{\gamma_{T,d}}{1+\lambda} - o(1)$.*

Proof. Recall that e_i denotes the number of E agents who arrived after time iT and are present in the pool at time $(i+1)T$ before the execution of the matching. By Lemma 19, for every integer $i \geq 0$ we have that $\mathbb{E}[e_i] = \gamma_{T,d}mT$. By Lemma 21, the matching executed at the matching time $(i+1)T$ involves at least e_i cross-edges whp. Therefore, the expected number of E agents matched in every executed matching is at least $\gamma_{T,d}mT(1-o(1))$, which is also a lower bound on the expected number of matched H agents. The following bounds thus hold for the match rates of E and H agents:

$$q_E^B(m) \geq \frac{1}{mT}\gamma_{T,d}mT(1-o(1)) = \gamma_{T,d}(1-o(1)),$$

$$q_H^B(m) \geq \frac{1}{(1+\lambda)mT}\gamma_{T,d}mT(1-o(1)) = \frac{1}{1+\lambda}\gamma_{T,d}(1-o(1)).$$ □

Proposition 10. *For a fixed batching policy with batch length T, the match rates of E agents and H agents as m grows large are $q_E^B = \gamma_{T,d}$ and $q_H^B = \frac{\gamma_{T,d}}{1+\lambda}$, respectively.*

Proof. For every agent type, Lemma 20 provided an upper bound for the match rate of agents of that type, and Lemma 22, provided a matching lower bound. The upper and lower bounds are $\gamma_{T,d}$ for E agents, and $\frac{\gamma_{T,d}}{1+\lambda}$ for H agents. This concludes the proof. □

Recall that $\gamma_{T,d} = \frac{1-e^{-T/d}}{T/d}$. The above proposition directly proves the claim of part ii of Proposition 3 about the match rates under the batching policy.

F.4. *Analysis of waiting time*

Lemma 23. *For every agent type $\Theta \in \{E,H\}$, $w^B_\Theta(m) = d(1-q^B_\Theta(m))$.*

Proof. The proof follows directly from Lemma 18. By that lemma, $q^B_\Theta(m) = 1 - \frac{w^B_\Theta(m)}{d}$. Rearranging the equality implies that $w^B_\Theta(m) = d(1-q^B_\Theta(m))$. □

G. PROOFS FOR PROPOSITION 3 AND THEOREM 1

Proof of Proposition 3 We analysed the match rate and waiting time under the greedy and batching policies respectively in Sections E and F. In particular, the claims about the match rate and waiting time under greedy policy (i.e. part (i) of the proposition) were proved in Sections E.1 and E.2, respectively. The claims about the match rate and waiting time under the batching policy (i.e. part (ii) of the proposition) were proved in Sections F.3 and F.4, respectively. The analysis of the patient policy is deferred to the Supplementary Appendix, Section v. The claims about the match rate and waiting time under the patient policy (i.e. part (iii) of the proposition) are proved there. □

Proof of Theorem 1 In Proposition 4, we showed that, under any policy, the match rate of hard-to-match agents is at most $\frac{1}{1+\lambda}$ and their expected waiting time is at least $\frac{\lambda d}{1+\lambda}$. On the other hand, in part (i) of Proposition 3, we showed that as m approaches infinity, the match rates of hard- and easy-to-match agents under the greedy policy approach $(q^G_H, q^G_E) = (\frac{1}{1+\lambda}, 1)$, respectively, and their expected waiting times approach $(w^G_H, w^G_E) = (\frac{\lambda d}{1+\lambda}, 0)$. This proves the first part of the theorem about the optimality of the greedy policy. It remains to show that the batching and patient policies are not asymptotically optimal.

Claim 9. *For every $T > 0$, $\frac{1-e^{-T/d}}{T/d} < 1$.*

Proof. For all $z > 0$, $1 - z < e^{-z}$. Setting $z = T/d$ and rearranging the terms proves the claim. □

Recall that by part (ii) of Proposition 3, as m grows large a batching policy with batch length $T > 0$ achieves match rates of $(q^B_H, q^B_E) = (\frac{1-e^{-T/d}}{(1+\lambda)T/d}, \frac{1-e^{-T/d}}{T/d})$, for hard- and easy-to-match agents, respectively. This fact, together with 9, implies that $q^B_H < \frac{1}{1+\lambda} = q^G_H$ and $q^B_E < 1 = q^G_E$. This proves the claim about the sub-optimality of the batching policy.

To show that the patient policy is not asymptotically optimal, we recall part (iii) of Proposition 3, which shows that the expected waiting time of hard-to-match pairs under the patient policy approaches $w^P_H = d$ as m grows large. On the other hand, under the greedy policy, the expected waiting time of hard-to-match agents approaches $w^G_H = \frac{\lambda d}{1+\lambda}$ as m grows large, which is strictly smaller than w^P_H. Therefore, the patient policy is not optimal. □

REFERENCES

AGARWAL, N., ASHLAGI, I., AZEVEDO, E., FEATHERSTONE, C. R. and KARADUMAN, Ö. (2019), "Market Failure in Kidney Exchange", *American Economic Review*, **109**, 4026–4070.

AKBARPOUR, M., COMBE, J., HE, Y., HILLER, V., SHIMER, R. and TERCIEUX, O. (2019), "Unpaired Kidney Exchange: Overcoming Double Coincidence of Wants without Money" (Working paper).

——————, LI, S. and GHARAN, S. O. (2020), "Thickness and Information in Dynamic Matching Markets", *Journal of Political Economy*, **128**, 783–815.

ANDERSON, R., ASHLAGI, I., GAMARNIK, D. and KANORIA, Y. (2017), "Efficient Dynamic Barter Exchange", *Operations Research*, **65**, 1446–1459.

AQUILINA, M., BUDISH, E. B. and O'NEILL, P. (2020), "Quantifying the High-Frequency Trading arms Race: A Simple New Methodology and Estimates", *Chicago Booth Research Paper* (20-16).

ASHLAGI, I., BURQ, M., JAILLET, P. and MANSHADI, V. (2016), "On Matching and Thickness in Heterogeneous Dynamic Markets", in *Proceedings of the 2016 ACM Conference on Economics and Computation* (ACM) 765–765.

——————, JAILLET, P. and MANSHADI, V. H. (2013), "Kidney Exchange in Dynamic Sparse Heterogenous Pools", in *Proceedings of the Fourteenth ACM Conference on Electronic Commerce* (ACM) 25–26.

ASHLAGI, I., BINGAMAN, A., BURQ, M., MANSHADI, V., GAMARNIK, D., MURPHEY, C., ROTH, A. E., MELCHER, M. L. and REES, M. A. (2018) "Effect of Match-Run Frequencies on the Number of Transplants and Waiting Times in Kidney Exchange", *American Journal of Transplantation*, **18**, 1177–1186.

——————, GILCHRIST, D. S., ROTH, A. E. and REES, M. A. (2011), "NEAD Chains in Transplantation", *American Journal of Transplantation*, **11**, 2780–2781.

——————, BURQ, M., DUTTA, C., JAILLET, P., SABERI, A. and SHOLLEY, C. (2019), "Edge Weighted Online Windowed Matching", in *Proceedings of the 2019 ACM Conference on Economics and Computation* (ACM) 729–742.

BACCARA, M., LEE, S. and YARIV, L. (2020), "Optimal dynamic matching", *Theoretical Economics*, **15**, 1221–1278.

BIRO, P., BURNAPP, L., BERNADETTE, H., HEMKE, A., JOHNSON, R., VAN DE, J., KLUNDERT, AND MANLOVE, D. (2017), *First Handbook of the COST Action CA15210: European Network for Collaboration on Kidney Exchange Programmes (ENCKEP)*.

BLANCHET, J. H., REIMAN, M. I., SHAH, V. and WEIN, L. M. (2020), "Asymptotically Optimal Control of a Centralized Dynamic Matching Market with General Utilities" *arXiv preprint arXiv:2002.03205*.

BRÉMAUD, P. (2013), *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, Vol. 31 (Springer Science & Business Media).

BRÉMAUD, P. (2017), *Discrete Probability Models and Methods*, Vol. 10 (Springer) 978–983.

CANONNE, C. (2019), "A Short Note on Poisson Tail Bounds".

DOVAL, L. (2014), "A Theory of Stability in Dynamic Matching Markets" (Technical Report, Mimeo).

FERRARI, P., WEIMAR, W., JOHNSON, R. J., LIM, W. H. and TINCKAM, K. J. (2014), "Kidney Paired Donation: Principles, Protocols and Programs", *Nephrology Dialysis Transplantation*, **30**, 1276–1285.

FRIEZE, A. and KAROŃSKI, M. (2015), *Introduction to Random Graphs* (Cambridge University Press).

GENTRY, S. E. and SEGEV, D. L. (2015) "The Best-Laid Schemes of Mice and Men Often Go Awry; How Should We Repair Them?", *American Journal of Transplantation*, **15**, 2539–2540.

HARCHOL-BALTER, M. (2013), *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*, 1st edn (New York: Cambridge University Press).

HELD, P. J., MCCORMICK, F., OJO, A. and ROBERTS, J. P. (2016), "A Cost-Benefit Analysis of Government Compensation of Kidney Donors", *American Journal of Transplantation*, **16**, 877–885.

KHALIL, H. K. (2009), "Lyapunov Stability", *Control Systems, Robotics and Automation–Volume XII: Nonlinear, Distributed, and Time Delay Systems-I.* 115.

LI, Z., LIEBERMAN, K., MACKE, W., CARRILLO, S., HO, C.-J., WELLEN, J. and DAS, S. (2019), "Incorporating Compatible Pairs in Kidney Exchange: A Dynamic Weighted Matching Model", in *Proceedings of the 2019 ACM Conference on Economics and Computation*, 349–367.

LITTLE, J. D. C. and GRAVES, S. C. (2008), "Little's Law", in *Building Intuition* (Springer) 81–100.

LIU, T. X., WAN, Z. and YANG, C. (2018), "The Efficiency of A Dynamic Decentralized Two-sided Matching Market" (Working Paper).

MERTIKOPOULOS, P., NAX, H. H. and PRADELSKI, B. (2020), "Quick or Cheap? Breaking Points in Dynamic Markets", in *Proceedings of the 21st ACM Conference on Economics and Computation*, 877–878.

NIKZAD, A., AKBARPOUR, M., REES, M. A. and ROTH, A. E. (2019), "Financing Transplants' Costs of the Poor: A Dynamic Model of Global Kidney Exchange".

NORRIS, J. R. (1997), *Markov Chains.* Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press).

ROTH, A. E., SÖNMEZ, T. and ÜNVER, M. U. (2004), "Kidney Exchange", *The Quarterly Journal of Economics*, **119**, 457–488.

——————, ——————, AND ——————, (2007), "Efficient Kidney Exchange: Coincidence of Wants in Markets with Compatibility-based Preferences", *The American Economic Review*, **97**, 828–851.

SAIDMAN, S. L., ROTH, A. E., SÖNMEZ, T., ÜNVER, M. U. and DELMONICO, F. L. (2006), "Increasing the Opportunity of Live Kidney Donation by Matching for Two-and Three-Way Exchanges", *Transplantation*, **81**, 773–782.

ÜNVER, M. U., (2010), "Dynamic Kidney Exchange", *Review of Economic Studies*, **77**, 372–414.

WEST, D. B. (2000), *Introduction to Graph Theory*, 2nd edn (Prentice Hall).