Online Appendix
Speculative Fever: Investor Contagion in the Housing Bubble

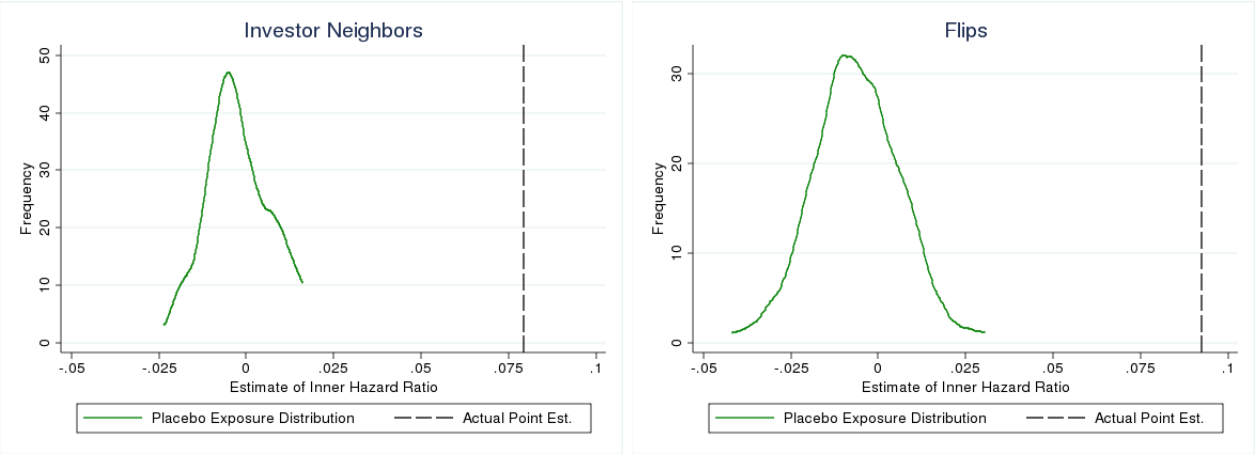
Patrick Bayer Kyle Mangum James W. Roberts

A Placebo Tests of Inner/Outer Research Design

As a falsification test of our main results, we calculate a false exposure measure by randomly drawing an at-risk tenure's inner ring (0.1 miles) exposure to both flips and investor neighbors conditional on their actual outer ring exposure. For example, if an at-risk tenure is exposed to 3 flips within 0.5 miles, we randomly draw how many of these are within 0.1 miles according the actual distribution of inner/outer exposure in the data. We then estimate comparable hazard regressions using this false-exposure data set.

In Figure A1, we report the distribution of the inner hazard effect from 100 draws of the false-exposure data set for each type of exposure. The placebo effect is centered tightly around zero. The results from actual data produces an estimate that is clearly to the right of the false exposure for both types of investing activity, indicating that homes actually exposed to very nearby investing activity are significantly more likely to become investors themselves. While this cannot identify the actual mechanisms at work–information sharing, word-of-mouth influence, etc.–the results of the placebo test are strong evidence of causal effects occurring within the narrow ring, city-block level of geography that induces homeowners to engage in investing activity.

Figure A1: Placebo Test of Contagion Mechanism: Full Data Sample with Placebo Exposure.



NOTES: The figure displays the distribution of the hazard estimator for the false (placebo) exposure to investment activity. The distribution comes from 100 random draws of placebo inner ring (0.1 mi.) exposures over the full data set. The dashed vertical line is the estimate from the full data sample (Table **??**).

B Robustness of Main Results

Table B1 illustrates the robustness of the results from the model of Table **??** to alternate clustering methods to account for spatial and temporal correlation in the model's error structure.

Table B2 illustrates that the baseline results in Table **??** with zip code-level fixed effects are robust to census tract-level fixed effects.

Table B3 presents a number of additional robustness checks related to sample selection. (Column 0 gives the baseline results from Table **??** above.) Section **??** described how we inferred whether the purchaser of the home was at-risk (i.e. used the home as their primary residence)[1] and whether an individual became an investor. We recognize the potential measurement error the inference on name matching entails, so we examine the sensitivity of the baseline results to our ability to identify at-risk tenures and flag investors in the data. Since this set of robustness checks changes the sample, we report the baseline hazard for the sample of at-risk individuals included in each alternative. All regressions use our research design, although for brevity we only report estimates of the innermost rings.

The first two columns consider the name matching algorithm used to infer investors. Column 1 uses only detailed names, i.e. those with middle names/initials and/or spouses listed, since these are less likely to be duplicated. The effect size as measured by hazard ratio is slightly smaller for flips, but larger for investor neighbors. Column 2 drops any names that are combinations of common names, defined as both first and last being in the top 20 percent of names observed in the data. Recall that we have already excluded any name with more than 42 properties attached, which removed common names like John Smith and Jose Lopez. This is an additional flag for a name like Michael Thompson, where Michael and Thompson

[1]Note that we also must observe the investor's primary residence to include this investor in the spatial match of the investor neighbor righthand-side variable. These sample selection checks refer to inclusion of the at-risk tenures (lefthand-side variable).

Table B1: Alternative Levels of Clustering Standard Errors.

	0	1	2	3	4	5	6	7	8
Investor Neighbor									
w/i. 0.1 mi	0.3228	0.3228	0.3228	0.3228	0.3228	0.3228	0.3228	0.3228	0.3228
	(0.0524)	(0.0535)	(0.0535)	(0.0540)	(0.0565)	(0.0605)	(0.0603)	(0.0576)	(0.0278)
w/i. 0.3 mi	0.0495	0.0495	0.0495	0.0495	0.0495	0.0495	0.0495	0.0495	0.0495
	(0.0268)	(0.0272)	(0.0274)	(0.0275)	(0.0304)	(0.0327)	(0.0343)	(0.0342)	(0.0227)
w/i. 0.5 mi	0.1685	0.1685	0.1685	0.1685	0.1685	0.1685	0.1685	0.1685	0.1685
	(0.0148)	(0.0151)	(0.0154)	(0.0162)	(0.0174)	(0.0199)	(0.0274)	(0.0330)	(0.0469)
Flip									
w/i. 0.1 mi	0.3769	0.3769	0.3769	0.3769	0.3769	0.3769	0.3769	0.3769	0.3769
	(0.0652)	(0.0657)	(0.0671)	(0.0676)	(0.0707)	(0.0599)	(0.0770)	(0.0931)	(0.1219)
w/i. 0.3 mi	0.0590	0.0590	0.0590	0.0590	0.0590	0.0590	0.0590	0.0590	0.0590
	(0.0330)	(0.0331)	(0.0333)	(0.0328)	(0.0353)	(0.0305)	(0.0282)	(0.0260)	(0.0241)
w/i. 0.5 mi	0.3159	0.3159	0.3159	0.3159	0.3159	0.3159	0.3159	0.3159	0.3159
	(0.0178)	(0.0179)	(0.0183)	(0.0189)	(0.0197)	(0.0211)	(0.0215)	(0.0288)	(0.0316)
Cluster level	Single	Single	Single	Single	Single	Double	Double	Double	Double
Cluster detail	Indiv. Tenure	Blk. Grp. X YQ	Tract X YQ	ZIP X YQ	Block Group	Indiv. Tenure Year-Mo.	Block Group Year-Qtr	ZIP code Year-Qtr	ZIP code Year
No. clusters 1	2,114,687	323,930	107,038	17,618	10,176	2,114,687	10,176	560	560
No. clusters 2						96	32	32	8

NOTES: The table displays regressions of Table **??**, column 2 (repeated in this table's column 1), with alternative clusterings of standard errors as noted.

Table B2: Baseline Results with Census Block Fixed Effects.

	1: Table **??**, Col 2	2: Table **??**, Col 5	3	4	5
Investor Neighbor					
w/i. 0.1 mi	0.3228	0.3462	0.3030	0.3012	0.3012
	(0.0524)	(0.0526)	(0.0525)	(0.0525)	(0.0525)
w/i. 0.3 mi	0.0495	0.0659	0.0427	0.0457	0.0464
	(0.0268)	(0.0269)	(0.0271)	(0.0271)	(0.0271)
w/i. 0.5 mi	0.1685	0.1479	0.1353	0.0616	0.0504
	(0.0148)	(0.0154)	(0.0158)	(0.0159)	(0.0159)
Flip					
w/i. 0.1 mi	0.3769	0.4110	0.4128	0.4048	0.4053
	(0.0652)	(0.0655)	(0.0661)	(0.0661)	(0.0661)
w/i. 0.3 mi	0.0590	0.0845	0.0869	0.0859	0.0858
	(0.0330)	(0.0332)	(0.0334)	(0.0334)	(0.0334)
w/i. 0.5 mi	0.3159	0.2033	0.2019	0.1181	0.1025
	(0.0178)	(0.0188)	(0.0191)	(0.0193)	(0.0194)
Constant	4.0720	4.3640	4.4608	4.1582	5.0148
	(0.0389)	(0.0461)	(0.0494)	(0.0768)	(0.3331)
Fixed Effects:					
Year				yes	
Year-Qtr		yes			yes
ZIP		yes			
Tract			yes	yes	yes

NOTES: The first two columns of the table are the baseline results from Table 5. The table illustrates that the baseline results are robust to the inclusion of finer geographic fixed effects, namely at the census tract level (for census tracts with very few sales, representing less than 5% of overall transactions, we aggregate to the zip code level). Standard errors in parentheses are clustered at the property tenure level. Coefficients have been multiplied by 10,000 for readability.

are common, but Michael Thompson was not so common as to already be removed.[2] This drops an additional 2 million monthly observations, with little effect on the results.

Columns 3 and 4 consider the geographic proximity of the investors' purchasing areas under the suspicion that a wide area may indicate two different individuals with the same (relatively uncommon) name. Column 3 drops any investor whose purchases are no closer than 50 miles, and column 4 drops anyone who purchases a property more than 50 miles from his/her other purchases. The loss of observations is small, indicating this is of low incidence, and the effect sizes are quite similar.

Our research design revolves around the "primary residence," which assumes the individual lives in the property we flag as such. The data contain two other sources of information (although also imperfect) on whether the property was actually owner-occupied. First, the HMDA data includes a flag for whether the loan application was for an owner-occupied home. Second, the assessor data match includes information on the owner's home mailing address; matching this to the property address gives another indicator for whether the home is considered owner-occupied.[3] Column 5 limits to individuals in properties flagged as owner-occupied. The results show that limiting our analysis to these tenures that meet these more stringent definition of an at-risk homeowner produces very similar results; to the extent we have misidentified primary residence, it does not appear to bias our results.

Overall, while we readily acknowledge the possibility of measurement error in our designation of investors, there is no evidence that it is driving our results. It would take a remarkable pattern of micro-level sorting by name similarity to randomly generate the patterns we observe.

[2]The commonality of first (and last) names was calculated ignoring the presence of middle names/initials or spouse names.

[3]Note that because the assessor data is overwritten each year by Dataquick, it reflects information from 2011. As a result, this measure or owner-occupancy can only be used for tenures that persist into the 2011 assessment year, limiting the number of observations for which this flag is useful primarily to those late in the sample.

Table B3: Robustness of Baseline Results to Sample Selection.

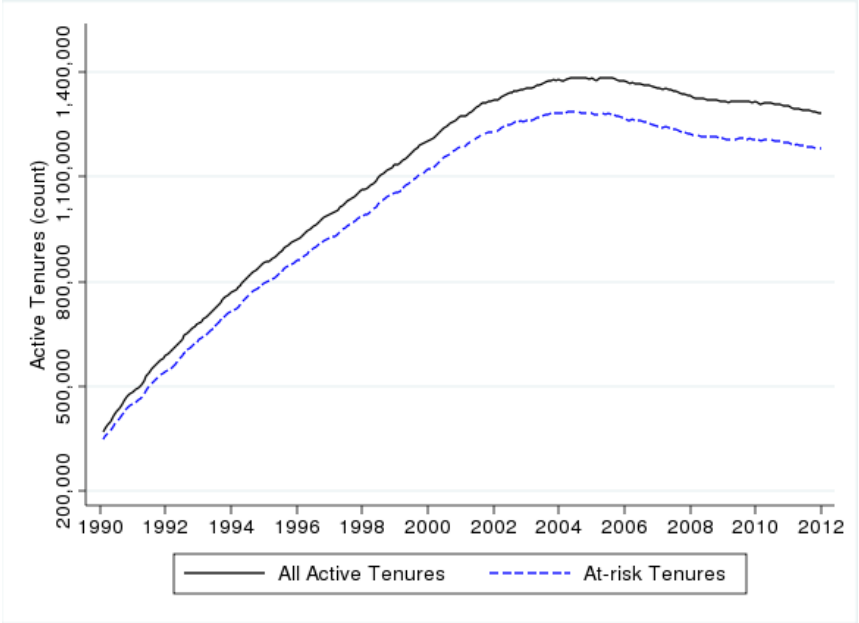
Specification:	0 Baseline	1 Detailed	2 Drop Common Names	3 Distant Property	4 Wide Purchase Area	5 Owner-Occupied
			Summary Statistics			
Tenures (N)	2,114,687	1,656,448	2,050,988	2,113,252	2,089,306	1,407,163
Obs (NT)	104,665,796	84,797,047	101,302,743	104,629,398	103,873,442	73,858,599
Entries	62,947	37,253	60,318	62,158	53,457	43,514
Entry Rate (x10,000)	6.01	4.39	5.95	5.94	5.15	5.89
			Regression Coefficients			
Investor Neighbor						
w/i. 0.1 mi	0.3228	0.3233	0.2849	0.2997	0.2906	0.3231
	(0.0524)	(0.0486)	(0.0494)	(0.0488)	(0.0458)	(0.0732)
Flip						
w/i. 0.1 mi	0.3769	0.2785	0.3496	0.3342	0.2648	0.3867
	(0.0652)	(0.0599)	(0.0613)	(0.0604)	(0.0563)	(0.0905)
			Marginal Effect to Hazard Rate			
Investor Neighbor						
w/i. 0.1 mi	0.0793	0.1051	0.0798	0.0850	0.0904	0.0877
	(0.0129)	(0.0156)	(0.0132)	(0.0130)	(0.0136)	(0.0184)
Flip						
w/i. 0.1 mi	0.0926	0.0815	0.0872	0.0847	0.0733	0.0932
	(0.0160)	(0.0192)	(0.0165)	(0.0162)	(0.0167)	(0.0228)

NOTES: The table presents various robustness checks for our main results. See the text for details and variable descriptions. Standard errors in parentheses are clustered at the property tenure level. Coefficients have been multiplied by 10,000 for readability.

C Descriptive Statistics: Main Estimation Sample

This appendix presents summary statistics describing basic features of the transaction database for Los Angeles, the primary metro area in our analysis.

Figure C1: Active Tenures and At-Risk Tenures Over Time, Monthly.



NOTES: The figure displays the count of active tenures and at-risk tenures (i.e. not, or not yet, investors) identified using the transaction data.

Table C1: Transaction and Property-level Summary Statistics.

	Los Angeles	
	Mean	Std. Dev.
Transactions	N=4,756,715	
Year of Transaction	1,999.5	6.7
Price ($)	266,919.4	200,618.0
Value ($ 2000)	213,124.7	141,536.4
Loan Present?	0.82	0.39
Equity < 5pct	0.27	0.44
	N=3,839,522	
LTV | Loan Present	0.86	0.15
	N=1,921,061	
Income*	102.6	139.7
	N=1,824,781	
Race: nonwhite*	0.49	0.50
Properties	N=2,271,384	
Year built	1,969.3	21.4
Sq. ft	1,657.3	646.8
No. beds	3.06	0.94
No. baths	2.16	0.78
Transactions	2.09	1.29

NOTES: The table shows transaction and property-level summary statistics for housing transactions data covering the five counties in the greater Los Angeles area (Los Angeles, Orange, Riverside, San Bernardino, and Ventura counties of California). Observation counts apply to subsequent rows within a column grouping.
Loan to value ratio (LTV) is measured relative to the price paid at the time of initial purchase. Value is the transaction price deflated by a metro-wide price index to year 2000 dollars. Property-level statistics are calculated over properties observed to transact during the sample period.
*For this variable to be observed, it must be present in the HMDA data and a reliable match to the transactions data had to be made.

Table C2: Transaction and Property-level Summary Statistics by Transaction Categorization, 2000-2007.

	1	2	3	4
	Non-investments	Investments	Investments, Investor Home ID'ed	Flips
			Mean (SD)	
Transactions				
N	1,429,900	280,855	181,351	52,779
Price ($)	355,983.40	347,893.10	356,927.60	296,939.60
	(223,978.30)	(218,759.70)	(215,931.90)	(199,726.70)
Value ($ 2000)	210,646.20	184,148.70	187,045.20	169,424.60
	(123,656.80)	(111,371.40)	(109,735.10)	(102,529.50)
Loan Present?	0.91	0.89	0.92	0.86
	(0.29)	(0.31)	(0.27)	(0.35)
Equity $<$ 5pct	0.33	0.36	0.38	0.30
	(0.47)	(0.48)	(0.49)	(0.46)
LTV	0.87	0.88	0.89	0.88
	(0.15)	(0.14)	(0.14)	(0.13)
(LTV N)	1,284,837	246,400	164,749	44,043
Properties				
N	1,101,216	255,307	172,426	50,268
Year built	1,971.40	1,968.34	1,968.42	1,964.85
	(21.82)	(22.54)	(22.38)	(23.26)
Sq. ft	1,649.91	1,535.77	1,534.49	1,469.17
	(653.85)	(615.27)	(614.09)	(581.50)
No. beds	3.04	2.96	2.97	2.90
	(0.94)	(0.95)	(0.95)	(0.93)
No. baths	2.17	2.05	2.05	1.98
	(0.77)	(0.77)	(0.77)	(0.76)
Transactions	2.58	3.08	3.03	4.04
	(1.43)	(1.61)	(1.61)	(1.62)

NOTES: The table shows transaction and property-level summary statistics for data that cover five counties in the Los Angeles area (Los Angeles, Orange, Riverside, San Bernardino, and Ventura) for properties flagged as non investments and several categories of investments. The sample is cleaned as described in main text. Loan to value ratio (LTV) is measured relative to the price paid at the time of initial purchase. Value is the transaction price deflated by a metro-wide price index to year 2000 dollars. Property-level statistics are calculated over properties observed to transact during the sample period.

D Descriptive Statistics: Additional Estimation Samples

This appendix presents summary statistics describing basic features of the transaction databases for the other two metro areas in our analysis, Boston and San Francisco.

Table D1: Transaction and Property-level Summary Statistics.

	San Francisco		Boston	
	Mean	Std. Dev.	Mean	Std. Dev.
Transactions	N=1,132,239		N=538,788	
Year of Transaction	1,999.7	6.8	1,999.9	6.8
Price ($)	404,618.2	282,122.6	258,774.6	199,066.2
Value ($ 2000)	408,710.0	246,616.8	255,488.7	177,518.7
Loan Present?	0.89	0.31	0.81	0.39
Equity < 5pct	0.16	0.36	0.11	0.31
	N=995,177		N=461,684	
LTV | Loan Present	0.81	0.16	0.78	0.17
	N=574,375		N=252,663	
Income*	124.3	124.6	94.7	114.2
	N=543,277		N=242,855	
Race: nonwhite*	0.43	0.49	0.14	0.35
Properties	N=604,575		N=332,757	
Year built	1,963.1	25.5	1,950.1	38.3
Sq. ft	1,672.2	674.6	1,782.0	787.8
No. beds	2.93	1.71	3.03	1.15
No. baths	2.01	1.26	1.88	0.94
Transactions	1.87	1.09	1.75	1.04

NOTES: The table shows transaction and property-level summary statistics for housing transactions data covering the metropolitan San Francisco Bay Area (Alameda, Contra Costa, Marin, San Mateo, and San Francisco counties of California) and metropolitan Boston (Essex, Middlesex, Norfolk, Plymouth, and Suffolk counties of Massachusetts). Observation counts apply to subsequent rows within a column grouping.
Loan to value ratio (LTV) is measured relative to the price paid at the time of initial purchase. Value is the transaction price deflated by a metro-wide price index to year 2000 dollars. Property-level statistics are calculated over properties observed to transact during the sample period.
*For this variable to be observed, it must be present in the HMDA data and a reliable match to the transactions data had to be made.

Table D2: San Francisco: Transaction and Property-level Summary Statistics by Transaction Categorization, 2000-2007.

	1	2	3	4
	Non-investments	Investments	Investments, Investor Home ID'ed	Flips
			Mean (SD)	
Transactions				
N	371,443	37,675	25,417	6,822
Price ($)	555,005.80	535,320.90	547,005.10	463,976.60
	(288,803.90)	(282,169.40)	(280,092.70)	(260,932.80)
Value ($ 2000)	390,991.60	355,740.20	362,913.50	319,002.20
	(196,098.10)	(187,117.10)	(185,832.70)	(170,190.60)
Loan Present?	0.94	0.89	0.91	0.86
	(0.23)	(0.32)	(0.28)	(0.35)
Equity $<$ 5pct	0.21	0.25	0.24	0.19
	(0.41)	(0.43)	(0.43)	(0.40)
LTV	0.83	0.84	0.84	0.84
	(0.16)	(0.15)	(0.15)	(0.14)
(LTV N)	344,734	32,857	22,882	5,736
Properties				
N	294,480	35,815	24,822	6,647
Year built	1,965.56	1,962.28	1,962.87	1,955.63
	(26.07)	(26.64)	(26.29)	(27.40)
Sq. ft	1,657.10	1,585.57	1,596.94	1,492.61
	(679.21)	(674.51)	(678.00)	(629.11)
No. beds	2.94	2.91	2.92	2.83
	(2.12)	(1.16)	(1.16)	(1.18)
No. baths	2.02	1.95	1.95	1.87
	(1.31)	(0.82)	(0.81)	(0.81)
Transactions	2.30	2.71	2.62	3.47
	(1.21)	(1.34)	(1.32)	(1.29)

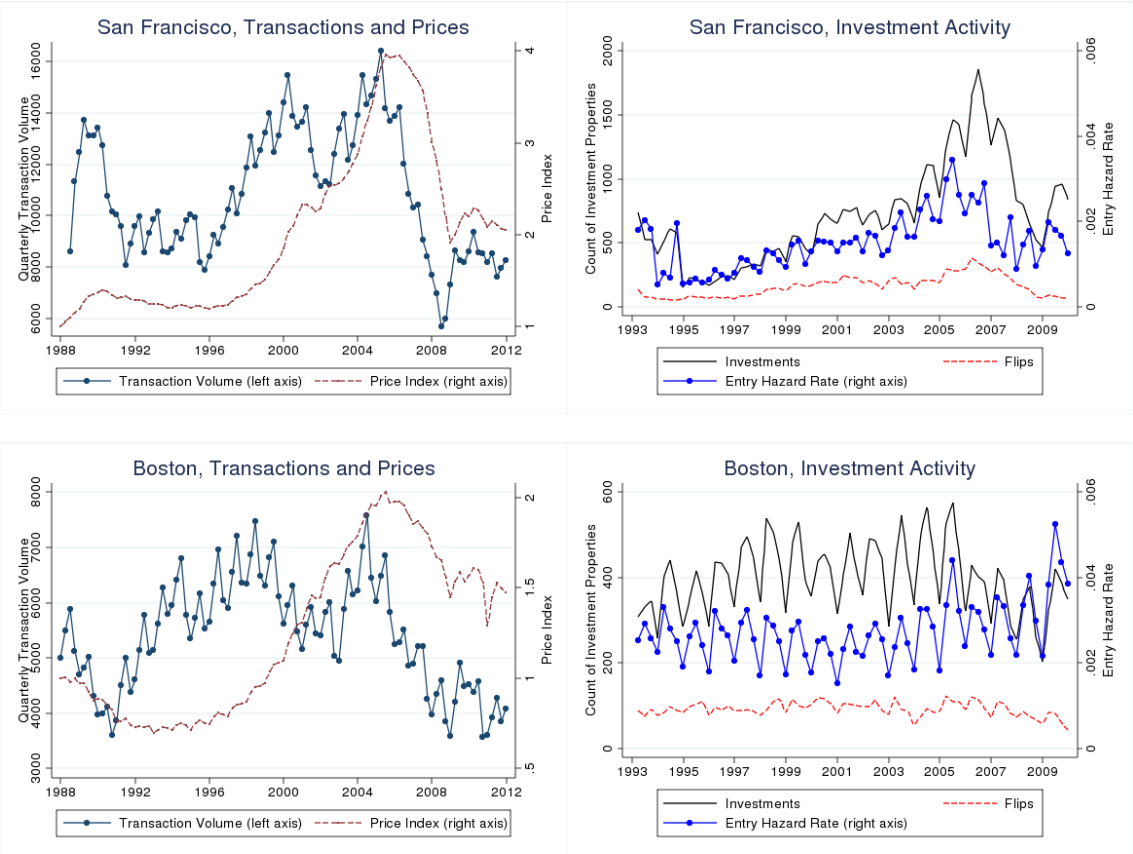
NOTES: The table shows transaction and property-level summary statistics for data that cover five counties in the San Francisco are (Alameda, Contra Costa, Marin, San Mateo, and San Francisco counties, California) for properties flagged as non investments and several categories of investments. The sample is cleaned as described in main text. Loan to value ratio (LTV) is measured relative to the price paid at the time of initial purchase. Value is the transaction price deflated by a metro-wide price index to year 2000 dollars. Property-level statistics are calculated over properties observed to transact during the sample period.

Table D3: Boston: Transaction and Property-level Summary Statistics by Transaction Categorization, 1998-2007.

	1	2	3	4
	Non-investments	Investments	Investments, Investor Home ID'ed	Flips
			Mean (SD)	
Transactions				
N	234,856	20,795	14,775	4,230
Price ($)	310,012.50	295,559.70	311,422.00	250,546.10
	(211,280.80)	(211,910.30)	(210,074.10)	(172,792.50)
Value ($ 2000)	248,416.70	234,728.00	247,550.90	199,694.20
	(165,504.40)	(165,226.80)	(164,497.40)	(135,498.10)
Loan Present?	0.85	0.76	0.83	0.66
	(0.36)	(0.43)	(0.38)	(0.47)
Equity $<$ 5pct	0.10	0.10	0.09	0.14
	(0.29)	(0.29)	(0.29)	(0.34)
LTV	0.78	0.77	0.77	0.81
	(0.17)	(0.16)	(0.16)	(0.14)
(LTV N)	196,489	15,118	11,933	2,444
Properties				
N	179,314	19,952	14,469	4,108
Year built	1,949.94	1,948.90	1,949.14	1,944.14
	(38.26)	(39.55)	(39.09)	(41.50)
Sq. ft	1,727.41	1,693.00	1,705.68	1,605.06
	(771.87)	(789.69)	(784.96)	(725.08)
No. beds	2.98	2.93	2.95	2.94
	(1.12)	(1.10)	(1.09)	(1.04)
No. baths	1.85	1.82	1.83	1.77
	(0.92)	(0.81)	(0.80)	(0.78)
Transactions	2.11	2.46	2.37	3.36
	(1.17)	(1.33)	(1.29)	(1.36)

NOTES: The table shows transaction and property-level summary statistics for data that cover five counties in the Boston area (Essex, Middlesex, Norfolk, Plymouth, and Suffolk counties, Massachusetts) for properties flagged as non investments and several categories of investments. The sample is cleaned as described in main text. Loan to value ratio (LTV) is measured relative to the price paid at the time of initial purchase. Value is the transaction price deflated by a metro-wide price index to year 2000 dollars. Property-level statistics are calculated over properties observed to transact during the sample period.

Figure D1: Additional Cities: Transactions, Prices, and Investment Acivity



NOTES: The figures give transaction, price and investment activity time series for the San Francisco and Boston metro areas. See text and the notes for Figures **??**, **??**, and **??** for additional details.

Table D4: Additional Estimation Sample Summary Statistics.

Metro Area:	San Francisco	Boston
At-risk tenures (N)	578,252	327,357
At-risk tenure-months (NT)	28,618,053	19,257,429
Entrants	13,294	6,577
Entry Rate (%)	2.30	2.01
Entry Hazard Rate (% x 10,000)	4.65	3.42
Investment Property Activity		
Mean Investments per Investor	1.36	1.26
Pct. Purchasing 1	78.40	81.62
Pct. Purchasing 2	14.46	13.97
Pct. Purchasing 3	4.08	2.83
Pct. Purchasing 4+	3.06	1.58

NOTES: The table reports summary statistics of the primary estimation sample for additional estimation samples of San Francisco, and Boston. *Entry hazard rate* is the outcome of interest. Additional statistics show the purchasing frequency of investors conditional on entry. Other definitions given in the main text.

Table D5: Additional Estimation Samples' Summary Statistics: Exposure to Investing Activity.

Panel A: San Francisco

Distance	Mean	Std. Dev.	Pct. w/. 0	Pct. w/. 1	Pct. w/. 2+
Investor Neighbors within:					
0.1	0.23	0.54	81.91	14.70	3.39
0.2	0.68	1.02	57.81	26.49	15.70
0.3	1.31	1.57	38.14	28.41	33.45
0.4	2.13	2.22	24.53	24.72	50.75
0.5	3.05	2.93	16.20	19.76	64.04
Flips within:					
0.1	0.10	0.34	91.38	7.62	1.00
0.2	0.31	0.66	77.09	17.42	5.49
0.3	0.60	1.00	62.32	24.09	13.59
0.4	0.99	1.41	49.12	27.07	23.81
0.5	1.42	1.86	38.90	27.16	33.94

Panel B: Boston

Distance	Mean	Std. Dev.	Pct. w/. 0	Pct. w/. 1	Pct. w/. 2+
Investor Neighbors within:					
0.1	0.14	0.43	88.83	9.43	1.75
0.2	0.34	0.69	74.90	18.98	6.12
0.3	0.63	1.00	60.08	25.45	14.47
0.4	1.00	1.35	47.18	27.80	25.02
0.5	1.35	1.67	38.43	27.43	34.14
Flips within:					
0.1	0.06	0.29	94.94	4.51	0.55
0.2	0.14	0.43	88.13	10.14	1.73
0.3	0.26	0.59	79.63	16.02	4.34
0.4	0.41	0.76	70.87	20.68	8.45
0.5	0.56	0.92	63.91	23.31	12.77

NOTES: The table reports summary statistics for investment exposures, the explanatory variable of interest, in the primary estimation sample for San Francisco and Boston metro areas. Definitions given in the main text. Spatial rings of exposure are inclusive of the narrower rings (e.g. 0.1 mile is also within 0.3 mile).

E Defining Neighbors Within a Building

A major challenge with conducting a within-building analysis is a data issue. In particular, while the data includes unit number, there is no information on the configuration of building(s) and, thus, no direct measure of the proximity of units within a building or way to tell whether two units are on the same floor. Thus, a primary challenge is to come up with a reliable way to understand whether two condo units within the same building were truly neighbors. This required understanding the labeling conventions of every multifamily complex in our dataset. To this end, we combined the use of an algorithm that could be automated to flag neighbors within a multifamily complex, and a brute force method of identifying the type of building structure (e.g. high rise versus townhomes) for the entire Los Angeles metropolitan area.

The idea is to make a dataset analogous to the surface distance neighborhood design in our baseline models, with the building functioning as the neighborhood and relatively closer property units within the building as being hyperlocal neighbors. We limit the sample to the set of larger multifamily buildings (20 units or more) in order to select properties that could have a meaningful amount of within-building variation, and we further restricted the sample to buildings constructed before the year 2000 to ensure that the buildings were occupied during our study period, and in particular, that the transactions we observe were not pre-sales of units in progress.

Among these properties, however, there is a substantial degree of variation in structure type, which complicates the assignment of within-building proximity. Some are condominium towers, separated by floors and served by elevators, while others are low-rise condominium communities spread out among clusters of apartment buildings. We often use the term "building" for simplicity, but technically the multifamily properties are a single street address–with a single observation of latitude and longitude coordinates, for which distance between units cannot be measured. There are multiple units at the address, but they are not necessarily housed in one building envelope.

Our first task is to classify buildings by structure type. We manually reviewed maps and satellite imagery (primarily using Google Maps and Streetview) to classify buildings (i.e., addresses) into four main types:[4] (1) low-rise apartment complexes comprised of one building of no more than three levels, (2) high-rise apartment towers of one building greater than three floors, (3) condominium communities, which are sets of multiple low rise apartment clusters with a single street address, and (4) townhome communities, similar to condominium communities but with the buildings arrayed as single family houses (often attached). Knowledge of the physical layout of the address provides guidance in assessing whether an arbitrary pair of units might be near one another according to the address's idiosyncratic numbering system.

Next, we turned to developing an algorithm for automating the designation of neighbors within the data. This required us to first decipher every building's unit numbering system. To do so we categorize each system into types based on the minimum and maximum number of digits observed in the unit number field (including alpha-numeric systems) to use in conjunction with the structure type information that we hand collected to distinguish "near

[4]We experimented with finer categories and obtained similar results. More important is the categorization of unit numbering systems and the interaction of this with broad structure type categories.

neighbors" from "far neighbors" within the building/address. For high-rise towers, this meant taking the unit number's digits in the hundreds and thousands places as group identifiers and digits in the ones and tens places for within-group. For example, units numbered 1101 to 1112 would comprise properties 1 to 12 in floor group 11. The majority of multifamily units, however, were in low rise or condominium communities, where the numbering systems were more often sequential (e.g., 1 to 300), or at least having many more units per floor. So when this occurs, we take the digits in the hundreds and tens places as the groups. For example, units 110 to 119 are properties 0 to 9 of group 11. Finally, after making the proximity designations, the data assembly procedure involves matching each at-risk homeowner tenure each month to entering investor neighbors and flip activity from the preceding year occurring within the building and within the nearby unit group.