Original research article
Evolution and Emerging Trends in HFT Research

ABSTRACT

Aims: In this paper, we try to study the evolution and emerging trends of High Frequency Trading (HFT) research by examining papers published in the Web of Science (WOS) between 1993 and 2017.
Study design: A total of 241 papers were included, and 1876 keywords from these articles were extracted and analyzed.
Place and Duration of Study: For tracing the dynamic changes of the HFT Research, the whole 24 year was further separated three consecutive periods: 1993-2002, 2003-2012, and 2013-2017.
Methodology: We used co-word analysis to reveal patterns and trends in the research by measuring the association strength of terms representative of relevant publications produced in HFT field. The Social Network Analysis (SNA) technique is adopted by using Ucinet to get keywords network, or knowledge network, to study the relationship of each research theme. NetDraw was applied to visualize network.
Results: Results indicate that HFT research has been strongly influenced by "market", "liquidity", "prices", "trades", "model", "stock", "stochastic", "statistics" and "finance", which represent some established research themes. They are major focuses and the bridges connecting to other research themes in HFT. "Market" revealed to be the most important keyword by betweenness centrality measuring for all three periods, and it has received consistent and high attention over the past three periods. "Stock" and "model" also received upward attention since 1993 until 2017. "Volatility" and "trades" were paid growing attention from 1993 through 2012 period, while 2013 to 2017 were not. "Liquidity", "finance" and "financial markets" are emerging theme since 2003 year due to they were not appeared in period of 1993 to 2002.
Conclusion: The above analysis provides an overview of HFT research and it suggests that market performance related keywords, which represent some established research themes, have become the major focus in HFT research. It also changes rapidly to embrace new themes. Especially, this research may make contribution to enlarge research method in that there is no co-word analysis research in HFT before.

Keywords: High Frequency Trading, HFT, co-word analysis, social network analysis, SNA, emerging trends

1. INTRODUCTION

As the stock market has become nearly exclusively electronic, advances in computer technology and automated algorithm trading have speeding the transmission and execution of security transaction orders, and thus establishing High Frequency Trading (HFT). HFT is an emerging, ever changing and rapidly evolving area with highly interdisciplinary in nature

for the markets, regulators, and the public. This diversity may root from the emerging nature of computing technology and its wide appeal as well as unique researcher and practitioner viewpoints. Many academics raised the controversy concerning about HFT. Even SEC Division of Trading and Markets Director Brett Redfearn admitted, “There are a lot of different definitions of HFT.” The diverse issues and findings in the field of HFT represent the introduction of ideas and even new concepts about HFT. What are the areas of focus in HFT? What are the developing trends in current research? Keywords have been generally identified as the words that reflect the research themes of individual publications that concern researchers. Further, network of keywords (co-word analysis) represents relationships of keywords among HFT papers. It is widely accepted that a higher co-occurrence frequency of two keywords in the literature indicates a closer relationship of these two themes. Two keywords occur in a same article is an indication of connection between the themes which they represent. Therefore, a comprehensive network perspective analysis is required to reveal the developmental trends or future orientation of possible new research field from HFT.

The co-word analysis is a comprehensive quantitative and visual analysis which was proposed as early as the late 70s in 20th century by French bibliometric scientists [1]. Co-word analysis can be used to analyze the knowledge structure and focus in a given field of research and thus visualize the relationships between various research themes [2]. Co-word analysis has been extensively used in literature-based research which include information retrieval, scientometrics, social science, psychological science, management science and medical research fields [3]. In this paper, our focus is to construct and analyze network of keywords (co-word analysis) by using the Social Network Analysis (SNA) techniques which have already been widely applied in many disciplines of science.

Specifically, this study will quantitatively analyze existing empirical and theoretical HFT papers to address the following objectives:

1) To construct network of keywords from HFT papers published in world leading journals during the period from 1993 to 2017.

2) To investigate the characteristics of network of keywords of HFT papers by utilizing Social Network Analysis (SNA) techniques.

3) To find and compare the change in network of keywords of HFT papers over time.

These investigations can help researchers to realize the breadth of HFT research and to establish future research directions and to provide an entry point to any academic, regardless of their prior knowledge of the theme.

2. METHODOLOGY

2.1 Publication search and keywords databases

The objective of the present work is to identify the important keywords from the scientific output on the latest advances in HFT, and to describe the characteristics of the network of keywords of HFT research. To achieve these goals, we selected the Web of Science (WOS), which includes SCIE and SSCI and A&HCI from the Institute of Scientific Information (ISI) Web of Science databases. WOS is the most important and frequently used source for a broad review of scientific accomplishment in all research fields. We constructed a database

composed of keywords from HFT papers published in the WOS during the 24-year period from 1993 to 2017. The keywords were obtained from following two sources: (1) Author Keywords and (2) Keywords Plus in the ISI database. Due to different words may represent same or similar ideas and concepts, we standardize the keywords before constructing the keyword network. The basic rule for the refinement of keywords was that all keywords with identical meaning or similar ideas or concepts or even misspelled keywords from different articles will be grouped and considered as a single keyword. This refinement leads to a meaningful keywords database.

2.2 Network of keywords by co-occurrence

Network of keywords by co-occurrence is composed of three continuous stages which include data extraction, data transformation, and data mapping. During the data extraction stage, core keywords are identified from HFT papers and are changed to a standard form. Then in the data transformation stage, co-word matrix is constructed by measuring the co-occurrence frequency of keywords in the articles. The co-occurrence matrix is then converted into a co-efficient matrix which reflects the degree of relations among keywords. At last in the mapping stage, the keywords are put into two dimensions on the co-efficient matrix. The final result is a keyword network where similar keywords are connected to each other.

2.3 Measurement

Network mapping is generally known as Social Network Analysis (SNA) technique which can be used to measure network centrality in the keyword network by calculating betweeness centrality, degree centrality or closeness centrality for each network quantitatively [4]. In order to understand the characteristics of the overall keyword network in HFT research, we selectively used betweenness centrality measuring. This is the extent to which a node lies on the paths between other nodes. It is measured as the fraction of the shortest paths between all pairs of other nodes in the network containing the node. In the keyword network, this represents the importance of a keyword in bridging subsets of keywords. A keyword that lies between two distinctive research themes can have high betweenness centrality even though it may have a small number of connections to other keywords in each theme [5]. For measurement, the Social Network Analysis (SNA) technique is adopted by using Ucinet to get keywords network, or knowledge network, to study the relationship of each research theme. NetDraw was applied to visualize network. It helps to obtain a clear sense of connectivity of keyword networks and to illustrate the overall patterns of networks over time. This method enables the researchers to explicitly understand representation of emerging themes.

3. RESULTS AND DISCUSSION

3.1 Occurrence and co-occurrence frequency of keywords

Keywords serve as an indicator of the importance of the research themes they represent. The keywords with higher frequency of both occurrence and co-occurrence can reflect research focuses to some extent in a special field [6]. The top ten keywords with higher frequency of occurrence are “market” (90), “liquidity” (79), “prices”(57), “trades” (56), “model” (53), “stock” (49), “stochastic” (43), “volatility” (42), “statistics” (38) and “finance”(37). The top ten keywords with higher frequency of co-occurrence are “market”, “prices”, “finance”, “liquidity”, “statistics”, “financial markets”, “stock”, “stochastic”, “model” and “trades” as

shown in Appendix A. After comparison, we can find that “market”, “liquidity”, “prices”, “trades”, “model”, “stock”, “stochastic”, “statistics” and “finance” not only have higher frequency of both occurrence and co-occurrence, but also indicate that these research themes are major focuses and the bridges connecting to other research themes in HFT. These findings show that these research themes attract more attention and have a closer relationship with other research themes in HFT. Notice that keywords like “High Frequency Trading” and “Algorithm(s)” have very broad meanings. Due to this kind of keywords are meaningless for this study, we excluded them from the above analysis.

3.2 Network of keywords by co-occurrence

Figure 1 shows network of keywords by co-occurrence (1993-2017). The nodes are the keywords. The size of nodes can reflect the frequency of keywords. Larger size of node means higher frequency of occurrence of keyword. The size and location of nodes depend on the total co-occurrence frequency of keywords. The lines between two nodes stand for the associations of two keywords, or represent the co-occurrence of these keywords in a paper. The thickness of line indicates the co-occurrence frequency of keyword pairs, or represents the number of times each pair of keywords was mentioned together in papers. The thickness of line is proportional to the closeness of connections between two keywords. The thicker line between two keywords, the closer their relationship is. The more co-occurrence between two keywords, the closer their relationship is. It shows the strength of the connection.

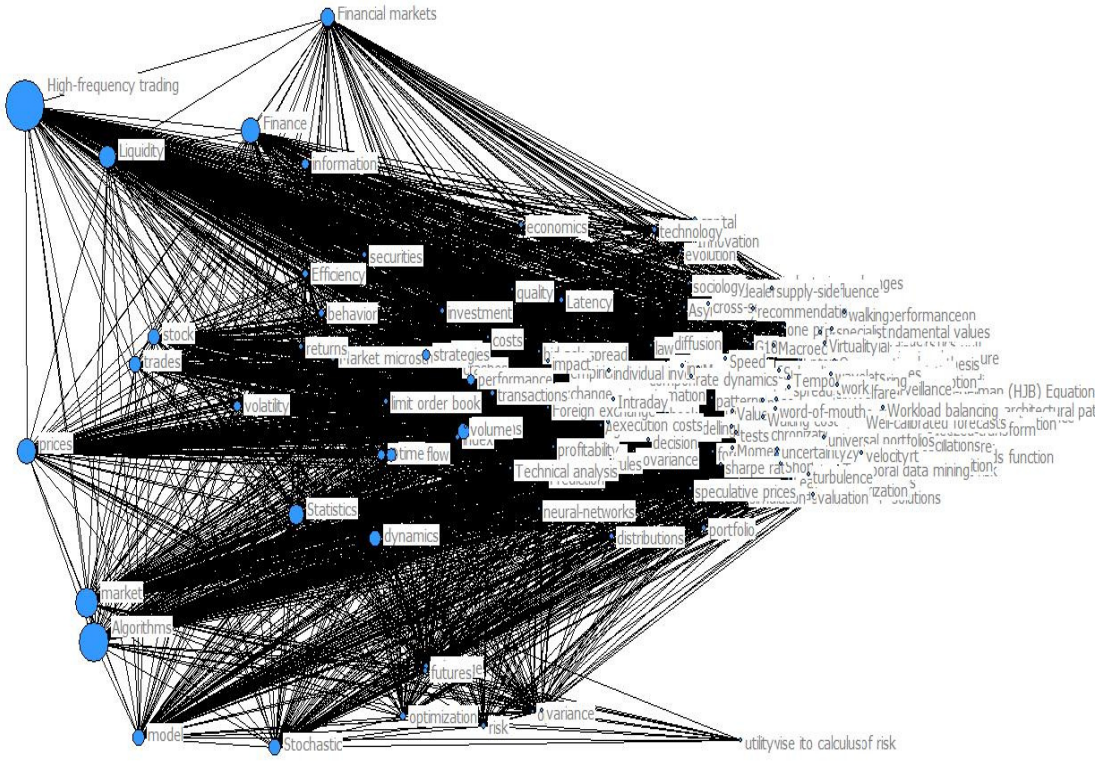


Fig.1. Network of keywords by co-occurrence (1993-2017)

According to Fig.1, keywords associated with market performance such as “financial markets”, “prices”, “market”, “model” and “stochastic” are located toward the periphery of the

network. In network analysis those keywords on the periphery might be considered as important research fronts, or emerging areas. Large nodes toward the centre of the network are more likely to represent important research paradigms. “Market”, “prices”, “finance”, “liquidity”, and “statistics” became important keywords in terms of Betweenness centrality as shown in Appendix A, which means that they have played an important role in bridging other research themes.

3.3 Changes in important keywords over time

How have the important keywords changed over time and what are the recent important keywords? In order to trace dynamic changes of the HFT Research, the whole 24 year was further separated three consecutive periods: 1993-2002, 2003-2012, and 2013-2017.We constructed three periods of network of keywords by co-occurrence, and compared the important keywords in the keyword networks constructed in order to thoroughly and precisely analyze the variations of trends. This comparison reveals some notable results. “Market” revealed to be the most important keyword by betweenness centrality measuring for all three periods, and it has received consistent and high attention over the past three periods. “Stock” and “model” also received upward attention since 1993 until 2017. “volatility” and “trades” were paid growing attention from 1993 through 2012 period, while 2013 to 2017 were not. The reason may be that “volatility” and “trades” are viewed as common sense already in HFT research until recent years. “Liquidity”, “finance” and “financial markets” are emerging theme since 2003 year due to they were not appeared in period of 1993 to 2002. The above analysis provides an overview of HFT research and suggests that market performance related keywords, which represent some established research themes, have become the major focus in HFT research. It also changes rapidly to embrace new themes. For the full list of Betweenness centrality measuring for these comparison, see Appendix A through Appendix D.

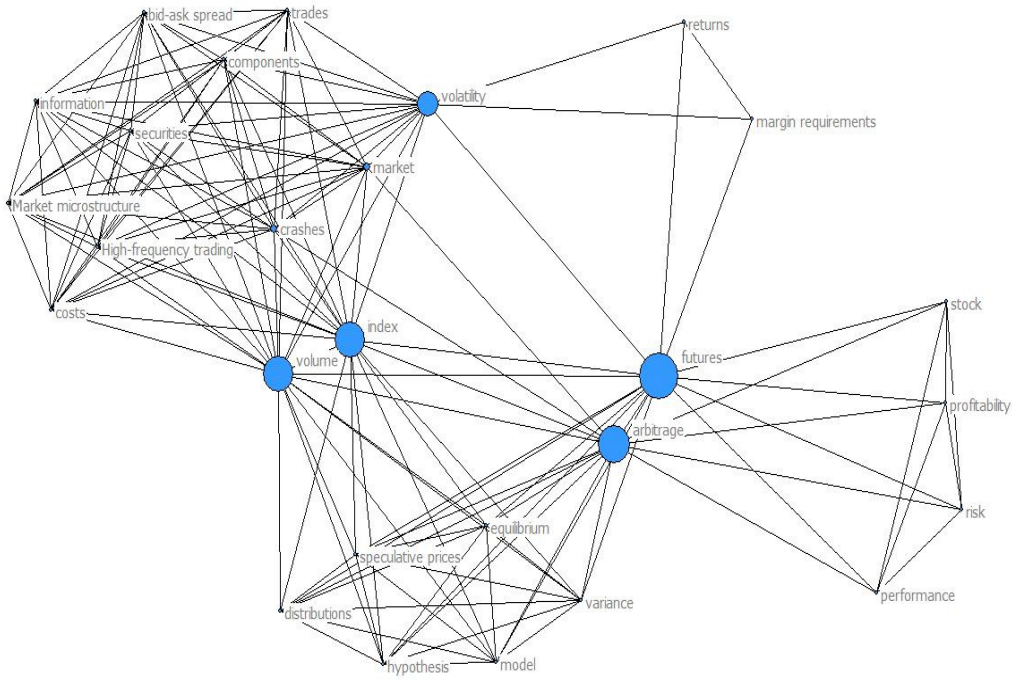
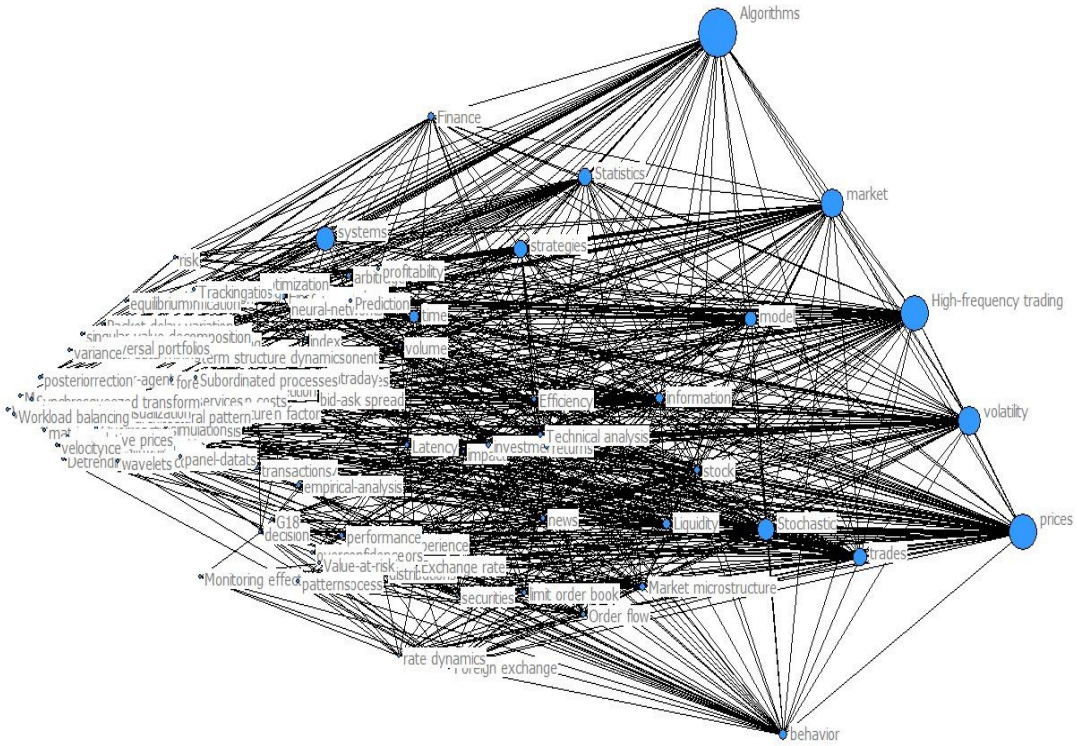
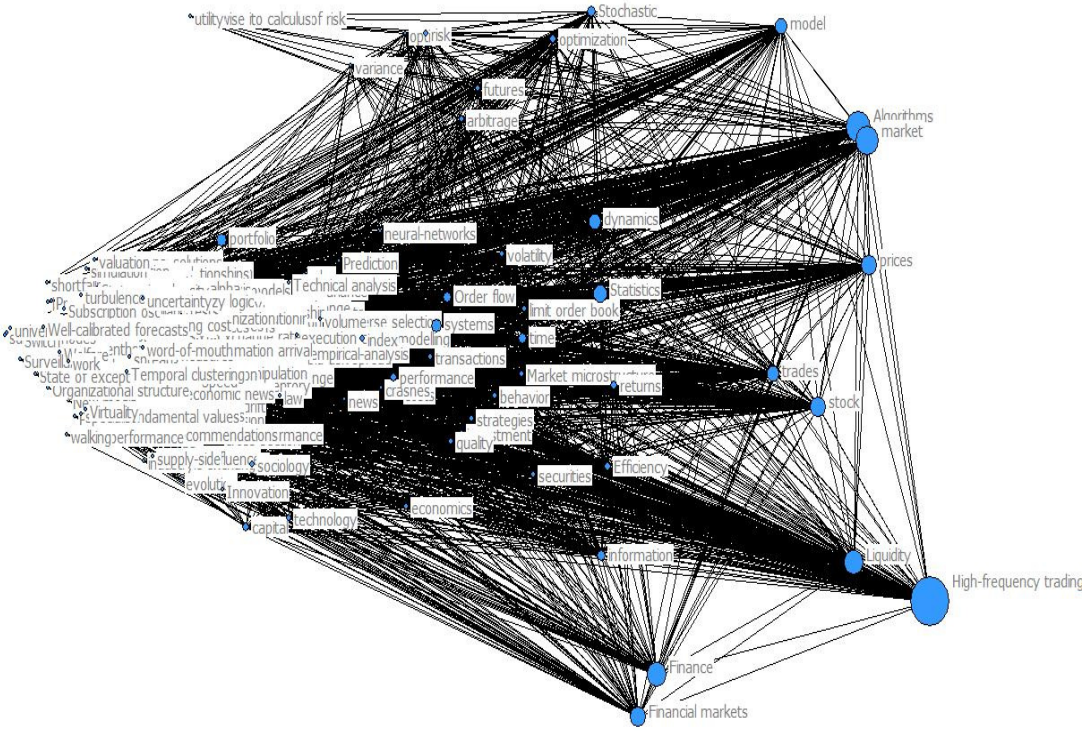


Fig.2. Network of keywords by co-occurrence (1993-2002)



182
183 Fig.3. Network of keywords by co-occurrence (2003-2012)
184



185
186 Fig.4. Network of keywords by co-occurrence (2013-2017)

4. CONCLUSION

In this article, we used co-word analysis which is a network-based text analysis and Social Network Analysis (SNA) technique which includes Ucinet and NetDraw to give a comprehensive understanding of HFT research during 1993 to 2017. We obtain some clear and reasonable results which can provide useful insights to better understand evolution and emerging trends in HFT research.

Results indicate that HFT research has been strongly influenced by “market”, “liquidity”, “prices”, “trades”, “model”, “stock”, “stochastic”, “statistics” and “finance”, which represent some established research themes. They are major focuses and the bridges connecting to other research themes in HFT. “Market” revealed to be the most important keyword by betweenness centrality measuring for all three periods, and it has received consistent and high attention over the past three periods. “Stock” and “model” also received upward attention since 1993 until 2017. “Volatility” and “trades” were paid growing attention from 1993 through 2012 period, while 2013 to 2017 were not. “Liquidity”, “finance” and “financial markets” are emerging theme since 2003 year due to they were not appeared in period of 1993 to 2002. The above analysis provides an overview of HFT research and it suggests that market performance related keywords, which represent some established research themes, have become the major focus in HFT research. It also changes rapidly to embrace new themes. Especially, this research may make contribution to enlarge research method in that there is no co-word analysis research in HFT before.

This study utilizes the advantage of the co-word analysis and such keywords analysis might be helpful to identify some fruitful future research opportunities. This research is just a preliminary and still has limitations need to be addressed. The Web of Science database does not completely cover the scientific research of HFT. In the future, comparative research with other method in the same HFT field could also be explored because different methods may have very different research emphases which would also be worthy of further exploration to extend HFT research theme.

REFERENCES

1. Callon, M., Rip, A., & Law, J. (Eds.). Mapping the dynamics of science and technology: Sociology of science in the real world. Springer; 1986.

2. Chen, Y., & Liu, Z. Y. The rise of mapping knowledge domain [J]. Studies In Science of Science, 2. 2005

3. Ravikumar, S., Agrahari, A., & Singh, S. N. Mapping the intellectual structure of scientometrics: A co-word analysis of the journal Scientometrics (2005–2010). Scientometrics. 2015;102(1), 929-955.

4. Larsen, E. N. An introduction to structural analysis: The network approach to social research: SD Berkowitz, Toronto: Butterworth; 1986.

5. Freeman, L. C. Centrality in social networks conceptual clarification. Social networks, 1978:1(3), 215-239.

6. Zong, Q. J., Shen, H. Z., Yuan, Q. J., Hu, X. W., Hou, Z. P., & Deng, S. G. Doctoral dissertations of Library and Information Science in China: A co-word analysis. Scientometrics, 2013;94(2), 781-799.

APPENDIX

Appendix A

Betweenness centrality measuring for all period (1993-2017)

		1993 - 2017				1993 - 2017	
No.	rank	Keywords		No.	rank	Keywords	
196	1	High-frequency trading	20972.666	99	46	costs	479.408
13	2	Algorithms	14873.707	142	47	empirical-analysis	428.283
258	3	market	11343.833	339	48	quality	425.657
333	4	prices	9020.214	435	49	volume	425.474
169	5	Finance	8998.353	430	50	variance	420.01
250	6	Liquidity	8372.251	378	51	sociology	403.265
389	7	Statistics	6550.62	329	52	power	381.97
170	8	Financial markets	6365.009	410	53	Technical analysis	378.11
391	9	stock	5479.472	175	54	Foreign exchange	343.948
390	10	Stochastic	5255.224	425	55	universal portfolios	339.003
278	11	model	5207.57	206	56	impact	305.744
419	12	trades	4817.368	158	57	Exchange rate	291.833
408	13	systems	4643.722	332	58	Prediction	284.327
132	14	dynamics	4483.216	301	59	options	283.819
416	15	time	3957.8	149	60	equilibrium	253.83
434	16	volatility	3471.744	78	61	competition	241.319
255	17	management	3040.512	10	62	Agent-based modelling	209.288
215	18	information	2945.935	216	63	Innovation	207.791
302	19	Order flow	2761.069	291	64	neural-networks	194.051
392	20	strategies	2683.609	213	65	individual investors	186.9
319	21	performance	2651.219	100	66	covariance	172.821
139	22	Efficiency	2355.799	91	67	Content-based	161.023
37	23	behavior	1895.172	39	68	bid-ask spread	158.613
259	24	Market microstructure	1890.566	115	69	decision	150.691
300	25	optimization	1622.698	371	70	sharpe ratio	140.984
352	26	returns	1461.738	155	71	evolution	129.907
210	27	index	1404.703	337	72	profitability	124.693
249	28	limit order book	1389.61	364	73	selection	123.284
54	29	capital	1313.427	256	74	Manipulation	117.729
326	30	portfolio	1312.679	422	75	turbulence	115.056
19	31	arbitrage	1276.097	159	76	execution costs	98.269
240	32	Latency	1247.931	242	77	law	94.131
184	33	futures	1177.844	360	78	rules	85.183
363	34	securities	1130.502	244	79	Lead-lag relationship	82.599
411	35	technology	1065.972	157	80	exchange	76.936
354	36	risk	1020.679	9	81	Adverse selection	70.177
226	37	investment	974.707	18	82	Approximation	68.299
138	38	economics	806.251	160	83	experience	63.362
293	39	news	750.328	22	84	ask	51.969
420	40	transactions	713.42	27	85	Asymmetry	40.406
30	41	Automation	681.092	70	86	Codings	40.025
122	42	diffusion	652.382	112	87	dealer	39.832
128	43	distributions	584.501	62	88	classification	39.252
101	44	crashes	532.398	382	89	speculative prices	38.483
297	45	Online learning	501.978	223	90	Intraday	38.163

Appendix B

Betweenness centrality measuring for the first sub-period (1993-2002)

		1993 - 2002	
No.	rank	Keywords	
8	1	futures	91.474
1	2	arbitrage	69.006
11	3	index	68.29
27	4	volume	68.29
26	5	volatility	46.118
5	6	crashes	9.371
14	7	market	9.371
2	8	bid-ask spread	0
3	9	components	0
4	10	costs	0
6	11	distributions	0
7	12	equilibrium	0
9	13	High-frequency trading	0
10	14	hypothesis	0
12	15	information	0
13	16	margin requirements	0
15	17	Market microstructure	0
16	18	model	0
17	19	performance	0
18	20	profitability	0
19	21	returns	0
20	22	risk	0
21	23	securities	0
22	24	speculative prices	0
23	25	stock	0
24	26	trades	0
25	27	variance	0

Appendix C

Betweenness centrality measuring for the second sub-period (2003-2012)

		2003 - 2012				2003 - 2012	
No.	rank	Keywords		No.	rank	Keywords	
5	1	Algorithms	2723.209	119	46	risk	21.541
68	2	High-frequency trading	1947.709	44	47	economics	21.067
113	3	prices	1943.382	62	48	futures	19.741
153	4	volatility	1592.97	13	49	bid-ask spread	18.193
86	5	market	1466.067	149	50	Value-at-risk	15.594
129	6	Stochastic	1124.416	75	51	Intraday	12.218
136	7	systems	1108.412	15	52	Boosting	2.133
144	8	trades	844.66	1	53	1st passage	0
128	9	Statistics	844.355	2	54	Active measurement	0
131	10	strategies	776.203	3	55	Adaptive trader-agents	0
92	11	model	714.488	4	56	Agent-based modelling	0
141	12	time	550.082	6	57	amorphous solids	0
84	13	Liquidity	393.819	7	58	anomalous diffusion	0
73	14	information	387.353	8	59	Approximation	0
12	15	behavior	379.06	10	60	Asynchronous data	0
55	16	Finance	330.972	14	61	Binary classification	0
87	17	Market microstructure	305.516	16	62	C33	0
130	18	stock	276.927	17	63	C41	0
41	19	distributions	198.327	18	64	C50	0
137	20	Technical analysis	190.978	19	65	cascades	0
95	21	news	186.846	20	66	choice	0
80	22	Latency	164.73	21	67	classification	0
9	23	arbitrage	163.332	22	68	Cloud computing	0
45	24	Efficiency	135.079	23	69	Codes of conduct	0
46	25	empirical-analysis	134.262	24	70	Codings	0
100	26	Order flow	130.235	25	71	Commodity hardware	0
58	27	Foreign exchange	128.547	26	72	Common factor	0
76	28	investment	128.284	27	73	Commonality	0
118	29	returns	114.102	28	74	competition	0
83	30	limit order book	109.02	29	75	component analysis	0
56	31	Financial markets	97.433	30	76	components	0
107	32	performance	87.288	31	77	continuous double auction	0
71	33	index	56.19	32	78	costs	0
115	34	profitability	56.01	33	79	covariance	0
36	35	decision	55.881	34	80	crashes	0
52	36	experience	54.663	35	81	Data stream processing	0
50	37	Exchange rate	52.486	37	82	Detrending	0
70	38	impact	50.131	38	83	diffusion	0
98	39	optimization	45.131	39	84	disposition	0
121	40	securities	34.477	40	85	Distributed processing	0
116	41	rate dynamics	31.575	42	86	dynamics	0
154	42	volume	30.481	43	87	EaaS	0
11	43	Automation	26.278	47	88	equilibrium	0
112	44	Prediction	26.052	48	89	error-correction	0
94	45	neural-networks	22.656	49	90	evolution	0

Appendix D

Betweenness centrality measuring for the third sub-period (2013-2017)

		2013 - 2017				2013 - 2017	
No.	rank	Keywords		No.	rank	Keywords	
163	1	High-frequency trading	15349.117	255	46	options	322.945
10	2	Algorithms	9033.231	277	47	power	320.466
217	3	market	8454.144	248	48	news	298.916
211	4	Liquidity	7070.412	365	49	volume	244.794
143	5	Finance	6733.11	362	50	variance	210.176
144	6	Financial markets	5455.364	246	51	neural-networks	206.152
280	7	prices	5301.453	7	52	Agent-based modelling	185.942
333	8	stock	5137.49	279	53	Prediction	181.464
331	9	Statistics	4706.375	62	54	competition	168.874
234	10	model	3940.322	171	55	impact	158.898
353	11	trades	3479.517	74	56	Content-based	144.998
111	12	dynamics	3275.315	309	57	selection	135.227
346	13	systems	2731.716	203	58	law	107.017
275	14	portfolio	2696.57	216	59	Manipulation	104.194
332	15	Stochastic	2543.021	120	60	empirical-analysis	97.351
215	16	management	2533.632	305	61	rules	95.875
256	17	Order flow	2276.406	83	62	covariance	85.314
352	18	time	2174.244	181	63	Innovation	85.282
180	19	information	2104.56	127	64	equilibrium	79.572
269	20	performance	1684.337	178	65	individual investors	77.577
254	21	optimization	1560.585	134	66	exchange	75.6
117	22	Efficiency	1499.502	33	67	bid-ask spread	74.359
44	23	capital	1204.612	136	68	execution costs	68.596
297	24	returns	1129.319	6	69	Adverse selection	64.104
31	25	behavior	1043.882	315	70	sharpe ratio	58.984
210	26	limit order book	981.42	348	71	Technical analysis	53.715
349	27	technology	929.715	17	72	ask	52.004
218	28	Market microstructure	885.26	135	73	Exchange rate	48.534
175	29	index	855.817	22	74	Asymmetry	45.24
364	30	volatility	833.319	139	75	facts	42.178
14	31	arbitrage	755.86	94	76	dealer	35.465
299	32	risk	752.568	38	77	book	33.13
308	33	securities	745.66	97	78	decision	31.873
191	34	investment	711.843	132	79	evolution	31.071
155	35	futures	635.551	192	80	issues	27.821
354	36	transactions	618.743	327	81	spread	23.048
334	37	strategies	590.264	106	82	discovery	22.219
202	38	Latency	579.702	133	83	Evolutionary computation	20.034
24	39	Automation	509.524	236	84	Momentum	18.852
103	40	diffusion	495.759	50	85	classification	17.174
116	41	economics	456.811	245	86	networks	16.995
285	42	quality	415.741	190	87	Inventory risk	16.062
321	43	sociology	409.809	8	88	aggressiveness	15.185
82	44	costs	401.178	123	89	entropy	14.797
84	45	crashes	334.673	89	90	cross-section	14.076

259

261 260