Supplemental Materials to "Emergence Scoring to Identify Frontier R&D Topics and Key Players"

Section A. Search Strategies

A1. Overview: The 4 Topics.
The search strategies reflect Boolean searches. To give a sense of these, we provide some details here. We note that exact replication is extremely challenging. The databases continue to update their content for years post-publication; search engines are modified over time; and most of these searches entail multiple strategies.

Two of the topics – DSSCs and NEDD – have roots in broader nanotechnology searches. We have searched within "nano" to develop initial datasets to jumpstart searching on nano sub-topics. Porter et al. (2008) and Arora et al. (2013) detail the search strategy refinement process. This entails several steps, notably:

- Download "nano*" items
  - Remove irrelevant records from that download (e.g., $NaNO_2$)
- Download 6 select other Boolean searches to capture nano-related topics
  - Apply two tiers of contingency relationships to boost precision (relevance to molecular scale activity).
- Retrieve articles published in select journals.

A2. NEDD Search
For NEDD, Zhou et al. (2014) detail the search strategy development and deployment, including how this was initiated in conjunction with the nano search. Excerpts here from that paper aim to provide the strategy and a sense of how it was implemented. See Zhou et al. (2014) for more complete details. [NOTE: The Tables, Figures, and Appendix are not copied here.]

In Table 1, we classify NEDD terms into seven main categories and several subgroups according to their functions. NEDD centers on "nano-enabled"—i.e., taking advantage of molecular scale properties of matter. Following a series of early search and retrieval trials, with topical expert review, the "N" (nano) portion of the search is broadened beyond explicit use of "nano*" terms, as seen in Appendix A.[10] Here we note, Figure 1 and Appendix A include all NEDD search terms in our current research. Such complexities pose a challenge to retrieve are representative set of NEDD research records. Our logic is to use terms in combination to operationalize "NEDD." We seek a high recall rate (i.e., capturing a high percentage of the relevant records pertaining to most important NEDD research topics), with reasonably good precision (i.e., by applying selection criterion to exclude undue noise).

To retrieve a suitable set of publications to characterize NEDD research activity, we focus on pharmaceutical/cargo (P)[1], nano-delivery-vehicle (N), characteristics of the delivery ap- proach (D), and drug cargo target (T), but largely set aside the B (biological processes), I (imaging), and H (helpers) categories. Including these would introduce high complexity and confounding issues (e.g., where to draw the line on imaging-related research?). Take the "T" category as an example—on one hand, searching for these terms will increase the search scope. On the other hand, incorporating particular diseases, like cancer, may distort the NEDD search results (by retrieving disproportionately more cancer-related publications).

After considerable probing and consultation, we key on two general search strategies: "D + P + N" and "D + T + N."
For test and evaluation, we searched variations of term combinations two ways: 1) directly from three databases: Web of Science ("WOS"; specifically its Science Citation Index Expanded), Medline, and Derwent Innovation Index ("DII"); 2) from a WOS nano data set compiled at Georgia Tech[10] that covers the "N" element with suitable term combinations.

This approach reflects a "Tech Mining" perspective that favors high recall at the expense of some precision. The logic is that later cleaning and text analyses can remove noise to conform to analytical aims. This process yielded the seven search strings indicated in Table 1-S. As detailed in Zhou et al. (2014), these are variously applied within the Georgia Tech nano search and/or the full external databases – Web of Science and MEDLINE, of interest presently.

Table 1-S. Search Phrases for the Base NEDD Downloads

1	TS=((deliver* or vehicle* or carrier* or vector* or "control* releas*") Near/4 (Drug* or pharmac))
2	TS=((deliver* or vehicle* or carrier* or vector* or "control* releas*" or transduct* or transfect* or transport* or translocat*) Near/4 agent*)
3	TS=((deliver* or vehicle* or carrier* or vector* or "control* releas*" or transfect*) Near/4 formulation*)
4	TS=((deliver* or vehicle* or carrier* or vector* or treat* or therap* or "control* releas*" or transduct* or transfect* or transport* or translocat*) Near/4 (siRNA or "short interfering RNA"))
5	TS= (deliver* or vehicle* or carrier* or vector* or treat* or therap* or "control* releas*" or transduct* or transfect* or transport* or translocat*) Near/4 (DNA or gene)
6	TS= (deliver* or vehicle* or carrier* or vector* or treat* or therap* or "control* releas*" or transduct* or transfect* or transport* or translocat*) Near/4 (Dox or Doxorubicin*)

7 TS=((deliver* or vehicle* or carrier* or vector* or treat* or therap* or "control* releas*"or transfect*) Near/4 ("RNA interference" or RNAi))

A3. DSSC Search

For **DSSCs**, Guo et al. (2012) note:

Boolean search phrasing is adjusted to suit particular database search configurations; here is the key: (((dye-sensitized or dye-sensitised or "dye sensitized" or "dye sensitised" or DSSC) WN KY) AND ((DSSC or "solar cell" or "solar cells" or photovoltaic or photovoltaics or photoelectrode or photoelectrochem* or photocurrent) WN KY)). We did discover a few other uses of "DSSC" (especially in telecommunications) and removed those records.

For the present analyses, Table 1 tallies records by time periods analyzed.

A4. Big Data Search

This presentation of the search strategy draws upon Huang et al. (2015). Figure 1 lays out the strategy. The paper details issues and our approach. Table 2-S summarizes the main search.

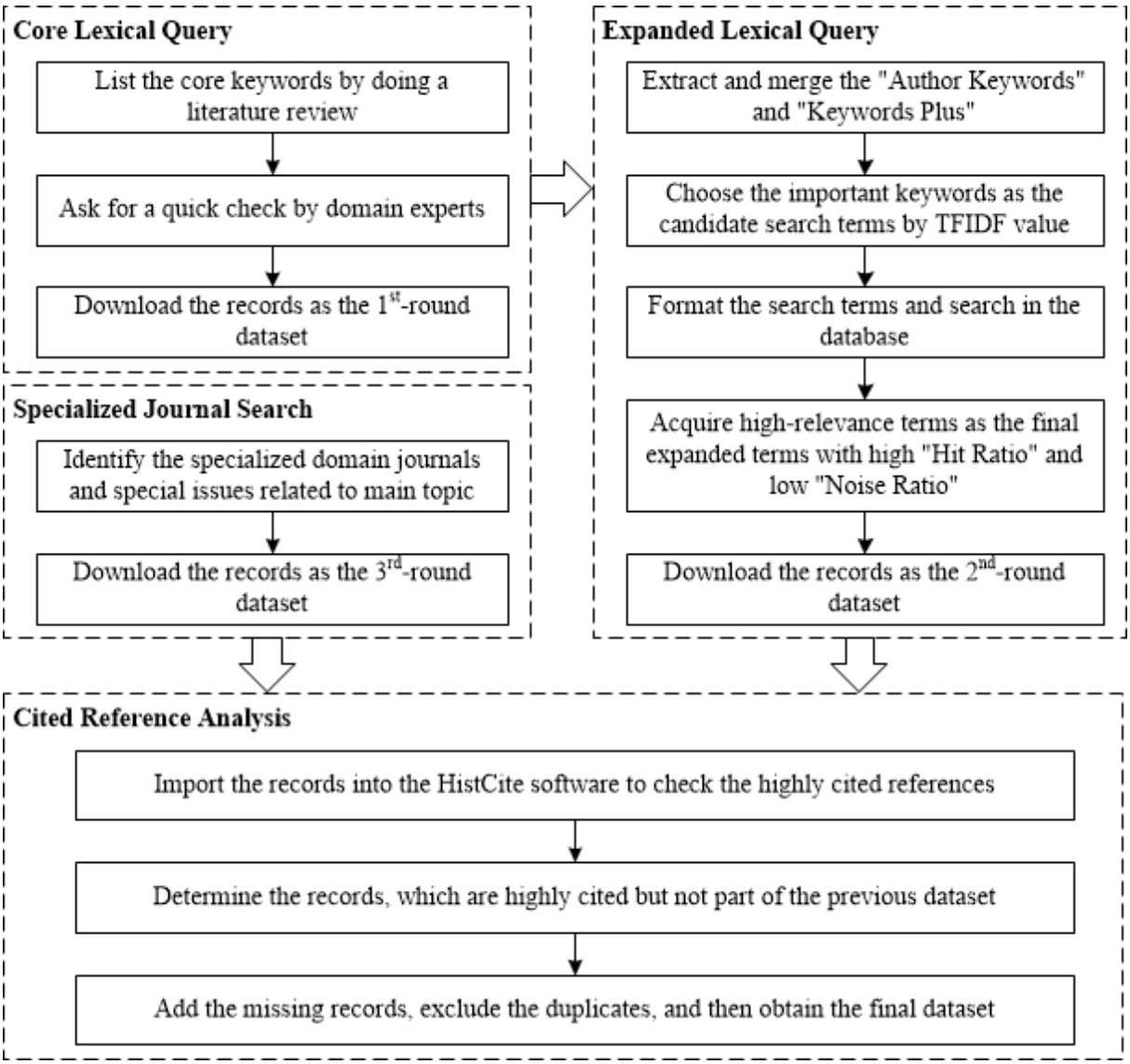

Figure 1. The framework of our search strategy for science

Table 2-S. The final search strategy.

No	Search Strategy	Search Terms
1	Core Lexical Query	TS= ("Big Data" or Bigdata or "Map Reduce" or MapReduce or Hadoop or Hbase or Nosql or Newsql)

2	Expanded Lexical Query	TS=((Big Near/1 Data or Huge Near/1 Data) or "Massive Data" or "Data Lake" or "Massive Information" or "Huge Information" or "Big Information" or "Large-scale Data" or Petabyte or Exabyte or Zettabyte or "Semi-Structured Data" or "Semistructured Data" or "Unstructured Data")
		TS=("Cloud Comput*" or "Data Min*" or "Analytic*" or "Privacy" or "Data Manag*" or "Social Media*" or "Machine Learning" or "Social Network*" or "Security" or "Twitter*" or "Predict*" or "Stream*" or "Architect*" or "Distributed Comput*" or "Business Intelligence" or "GPU" or "Innovat*" or "GIS" or "Real-Time" or "Sensor Network*" or "Smart Grid*" or "Complex Network*" or "Genomics" or "Parallel Comput*" or "Support Vector Machine" or "SVM" or "Distributed" or "Scalab*" or "Time Serie*" or "Data Science" or "Informatics*" or "OLAP")
3	Specialized Journals	The papers published in these specialized journals are not indexed by WOS
4	Cited Reference	The publications, which were cited more than 20 times, did not fulfill the criteria for inclusion (see paragraph "Cited Reference Analysis")

A5. Non-Linear Programming Search

This search algorithm was relatively straightforward, focusing on the phrase, "non-linear programming." The Web of Science search checked record relevance between use of keyword fields (Keywords-Author and/or Keywords-Plus) versus title and abstract Natural Language Processing (NLP) phrases. Review by two Georgia Tech faculty found the quality of emergent terms comparable from both.

A6. Search References

Arora, S.K., Porter, A.L., Youtie, J., and Shapira, P. (2013), Capturing new developments in an emerging technology: An updated search strategy for identifying nanotechnology research outputs, *Scientometrics*, 95 (1), 351-270. DOI: 10.1007/s11192-012-0903-6.

Y. Guo, C. Xu, L. Huang, A.L. Porter, "Empirically informing a technology delivery system model for an emerging technology: Illustrated for dye-sensitized solar cells, *R&D Management*, 42 (2) (2012) 133-149.

Porter, A.L., Youtie, J., Shapira, P., and Schoeneck, D.J., Refining Search Terms for Nanotechnology, *Journal of Nanoparticle Research*, Vol. 10 (5), 715-728, 2008.
[Online First: 10.1007/s11051-007-9266-y].

X. Zhou, A.L. Porter, D.K.R. Robinson, M.S. Shim, Y. Guo, Nano-enabled drug delivery: A research profile, *Nanomedicine: Nanotechnology, Biology and Medicine.* 10 (5) (2014) 889-896; *http://dx.doi.org/10.1016/j.nano.2014.03.001*.

Figure 1-A. High ESc5 Terms scoring above 1.77 with record counts for the 7-year active period

	# Records	Non-Linear Terms	>1.77 ESc5		# Records	Big Data Terms
1	106	mixed integer	5.261	1	622	big data
2	25	operating cost	4.304	2	119	data analytics
3	18	Mixed Integer Non-Linear Program MINLP	4.081	3	600	MapReduce
4	17	linear behavior	3.968	4	80	big data analytics
5	22	novel approach	3.257	5	436	Hadoop
6	12	model results	3.24	6	73	social media
7	16	non-linear function	3.123	7	61	Big Data process
8	7	mixed integer linear program MILP	2.933	8	151	framework MapReduce
9	15	non-linear behavior	2.8	9	130	social network
10	16	scheduling problem	2.747	10	66	Hadoop cluster
11	8	non-linear response	2.693	11	128	high performance
12	45	mixed integer non-linear program	2.671	12	175	large amount
13	10	computational experiment	2.66	13	50	big data application
14	17	heuristic algorithm	2.625	14	110	distributed file system
15	73	mixed integer non-linear	2.581	15	185	data process
16	12	present work	2.556	16	63	cloud environment
17	50	integer non-linear program	2.292	17	47	NoSQL database

18	8	fuel cost	2.247
19	8	manufacturing system	2.244
20	7	useful tool	2.134
21	42	production	2.096
22	19	MINLP problem	2.07
23	10	Mixed Integer Non Linear Programming MINLP problem	2.021
24	8	planning problem	2.011
25	9	real data	2.004
26	10	annual cost	1.905
27	9	calculated results	1.84
28	7	non-linear term	1.828
29	35	total cost	1.801

18	146	file system	
19	48	extensive experience	
20	65	hadoop MapReduce	
21	67	Hadoop distributed file system	
22	57	File System HDFS	
23	48	Apache Hadoop	
24	54	distributed file system HDFS	
25	68	HBase	
26	50	Hadoop Distributed File System	
27	53	MapReduce job	
28	35	cloud platform	
29	145	recent year	
30	27	private Cloud	
31	127	organizers	
32	47	improved performance	
33	44	Mapreduce application	
34	35	public cloud	
35	51	analytics	
36	56	process large-scale data	

	# Records	Dye-Sensitized Solar Cell Terms	ESc5
1	179	power conversion	13.063
2	174	power conversion efficiency	11.978
3	94	organic dye	9.278
4	121	electrochemical impedance	8.495
5	197	photovoltaic performance	8.185
6	128	electron microscopy	7.07
7	68	TiO(2)	7.066
8	51	extinction coefficient	6.714
9	46	TiO(2) film	6.428
10	71	density functional theory	6.196
11	54	solar cell application	6.174
12	51	TiO2 nanotube	6.16
13	35	dye sensitized solar cell application	5.657
14	149	surface area	5.542
15	85	glass substrate	5.466
16	148	efficient conversion	5.338
17	126	impedance spectroscopy	5.033
18	291	open circuit voltage	4.966
19	95	electrochemical impedance spectroscopy	4.742
20	74	tin oxide	4.211
21	34	molar extinction coefficient	3.917
22	81	X-ray diffraction	3.863
23	30	nanotube array	3.839
24	82	photovoltaic property	3.727
25	22	TiO2 nanotube array	3.686
26	22	organic sensitizer	3.681
27	31	dye N719	3.635
28	106	overall conversion efficiency	3.61
29	152	fabricated	3.563
30	40	X ray diffraction XRD	3.526
31	27	electron microscopy SEM	3.493
32	80	electron lifetime	3.334
33	41	density functional theory DFT	3.327

41	20	cyanoacrylic acid
42	48	transfer resistance
43	62	high efficiency
44	137	charge transfer
45	38	diffusion length
46	37	electrochemical impedance spectroscopy EIS
47	41	optical property
48	76	cyclic voltammetry
49	17	TiO(2) nanoparticles
50	22	charge collection
51	36	high performance
52	28	photocatalytic activity
53	20	DSSC application
54	23	field emission
55	26	surface morphology
56	19	nanocrystalline TiO(2) film
57	56	dye adsorption
58	17	quasi solid state DSSC
59	14	nanocrystalline TiO(2)
60	25	overall light
61	17	electrophoretic deposition
62	45	photoelectric conversion efficien
63	115	polymer electrolyte
64	15	electrochemical impedance spec
65	48	particle size
66	71	solar cell DSSC
67	28	hydrothermal method
68	31	anode
69	20	ethylene glycol
70	24	molecular structure
71	134	counter electrode
72	47	transmission electron microscop
73	17	high molar extinction coefficien

34	28	microscopy SEM
35	75	short circuit current
36	75	circuit current
37	22	ZnO nanorod
38	38	impedance spectroscopy EIS
39	45	charge transfer resistance
40	42	sensitized solar cell DSSC

3.247	74	94 efficiency eta
3.221	75	40 low cost
3.221	76	18 electron diffusion length
3.039	77	51 ionic liquid electrolyte
3.004	78	12 nanostructure ZnO
2.996	79	34 fluorine doped tin oxide
2.985	80	415 open circuit
	81	23 enhanced performance

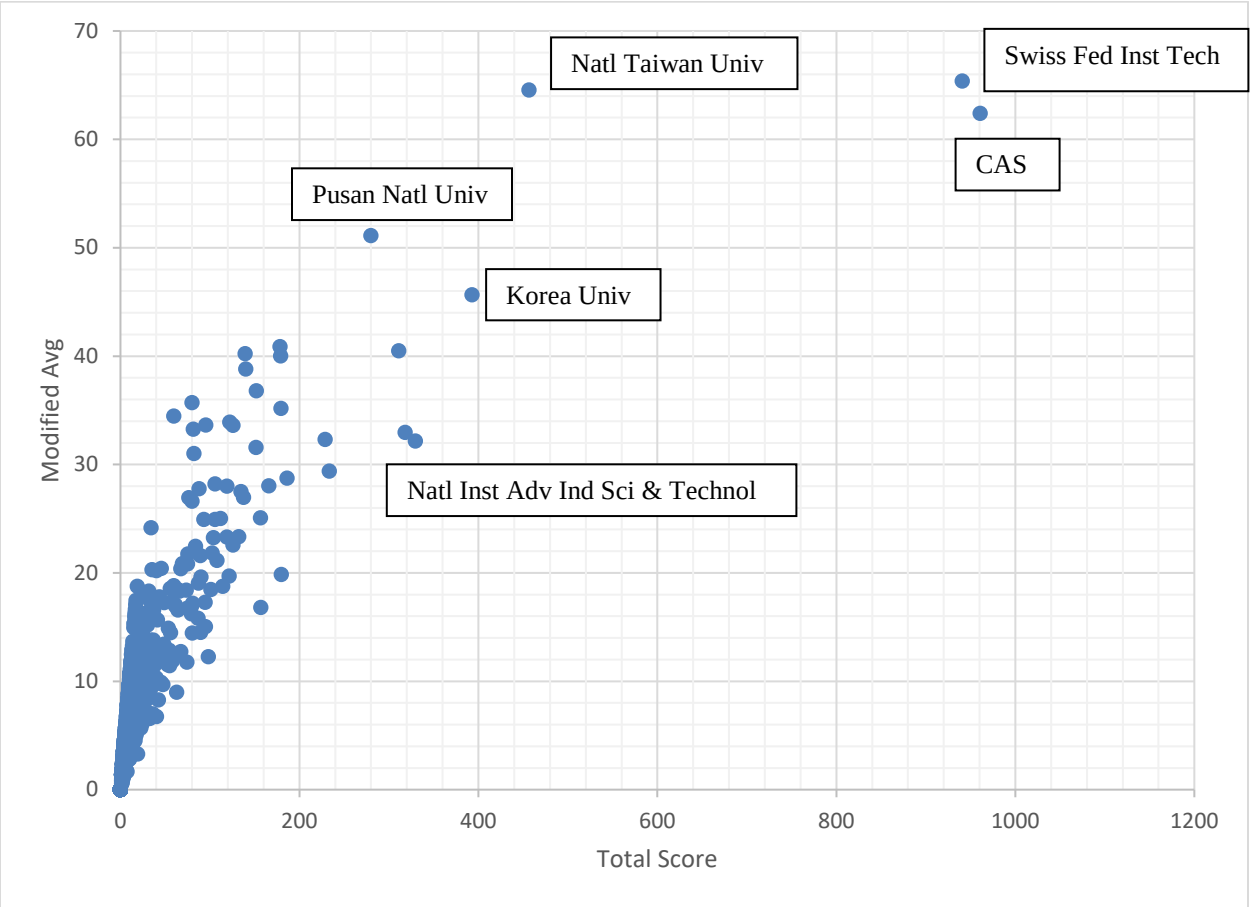

Figure 2-A. Normalized vs. Total EScores for Leading Organizations Publishing on DSSCs

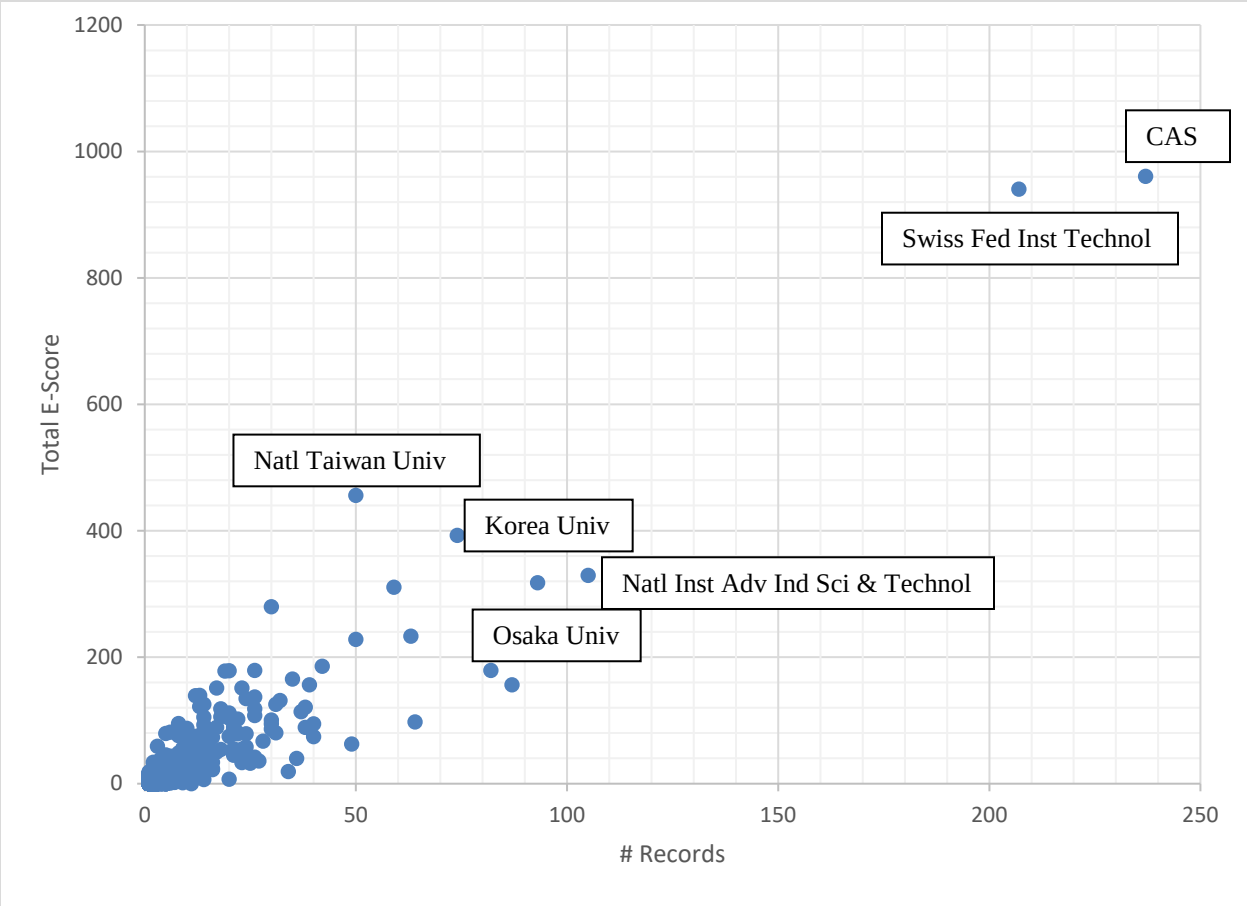

Figure 3-A. Total EScores for Leading Organizations Publishing on DSSCs by Number of Publications

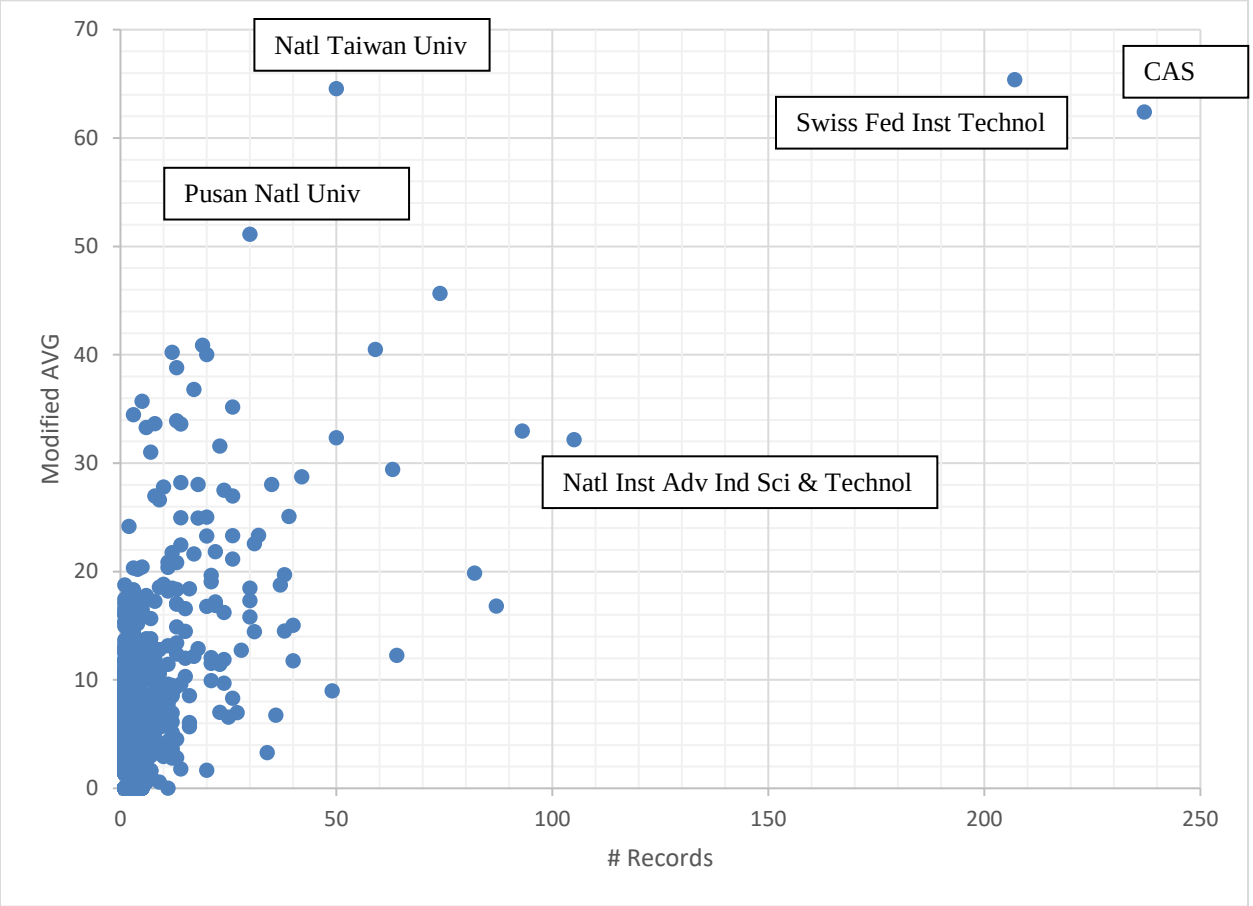

Figure 4-A. Normalized EScores for Leading Organizations Publishing on DSSCs by Number of Publications

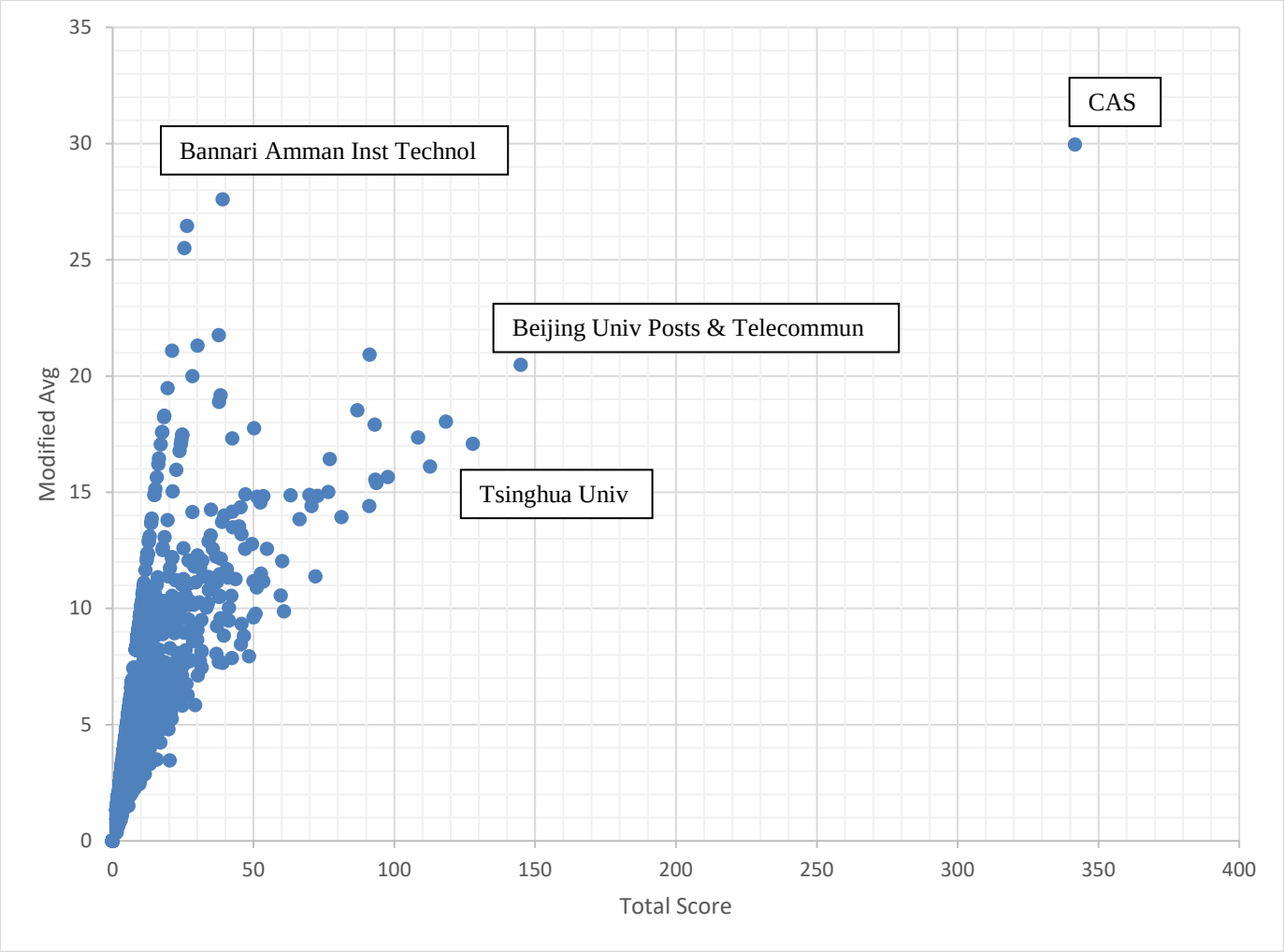

Figure 5-A. Normalized vs. Total EScores for Leading Organizations Publishing on Big Data

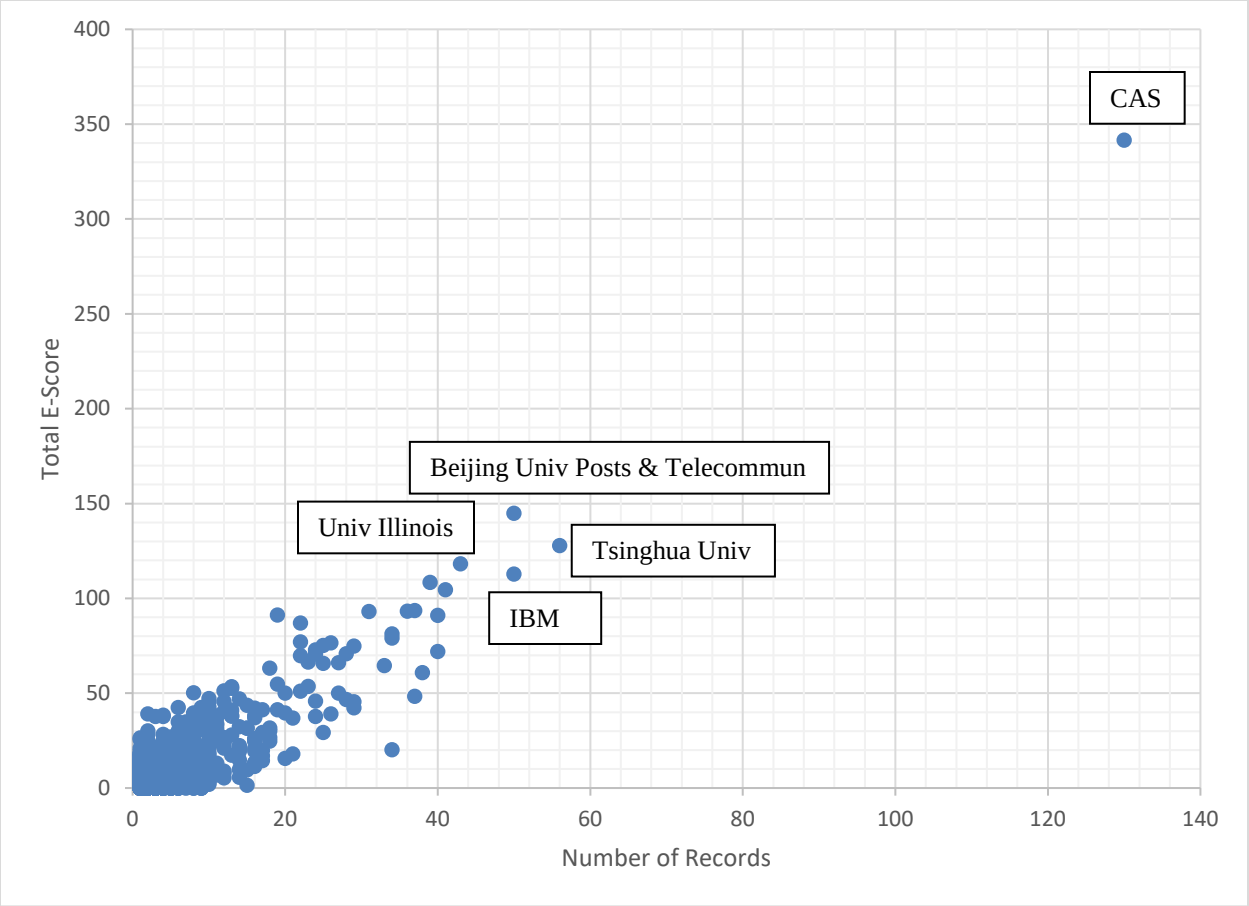

Figure 6-A. Total EScores for Leading Organizations Publishing on Big Data by Number of Publications

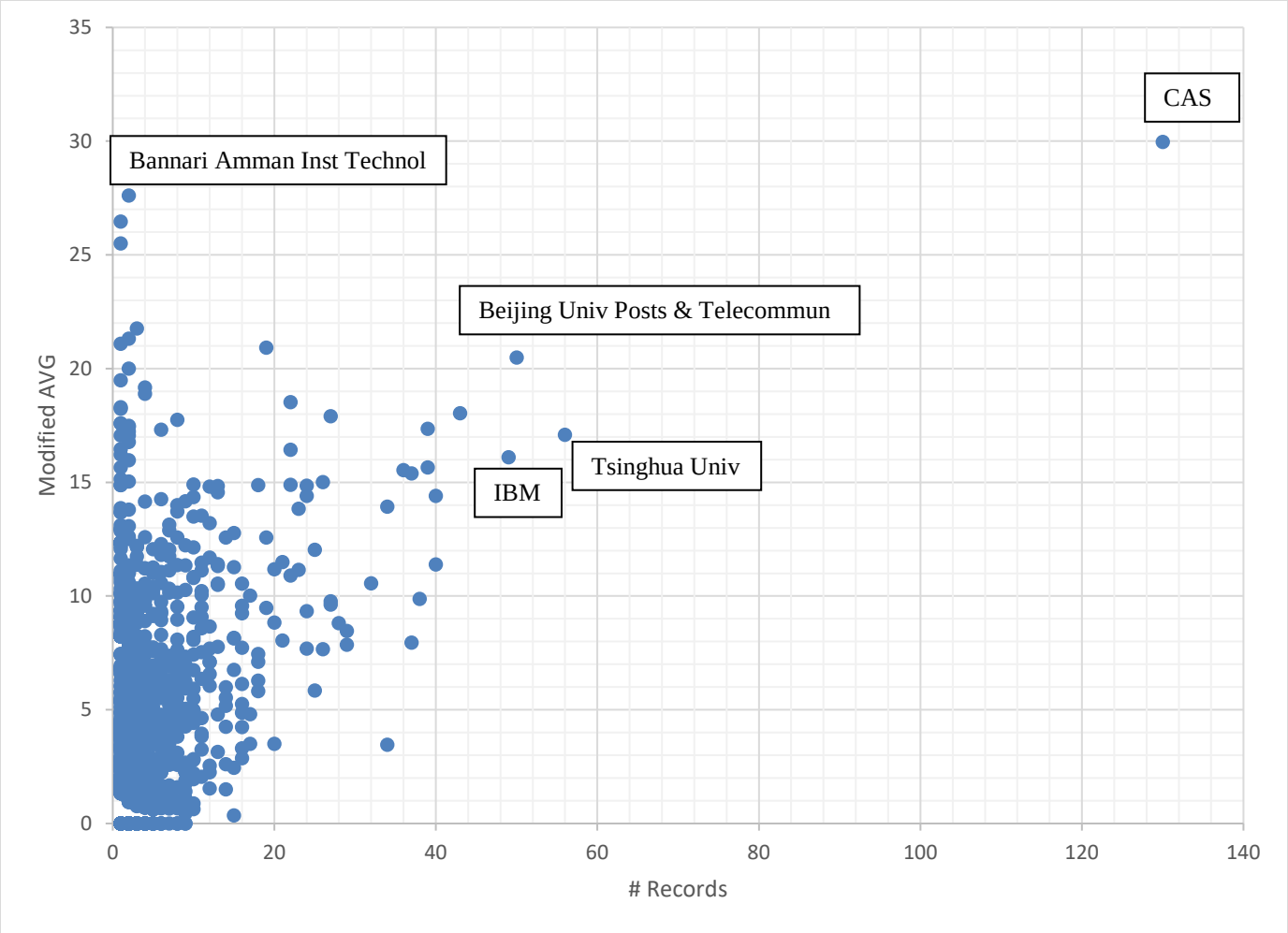

Figure 7-A. Normalized EScores for Leading Organizations Publishing on Big Data by Number of Publications

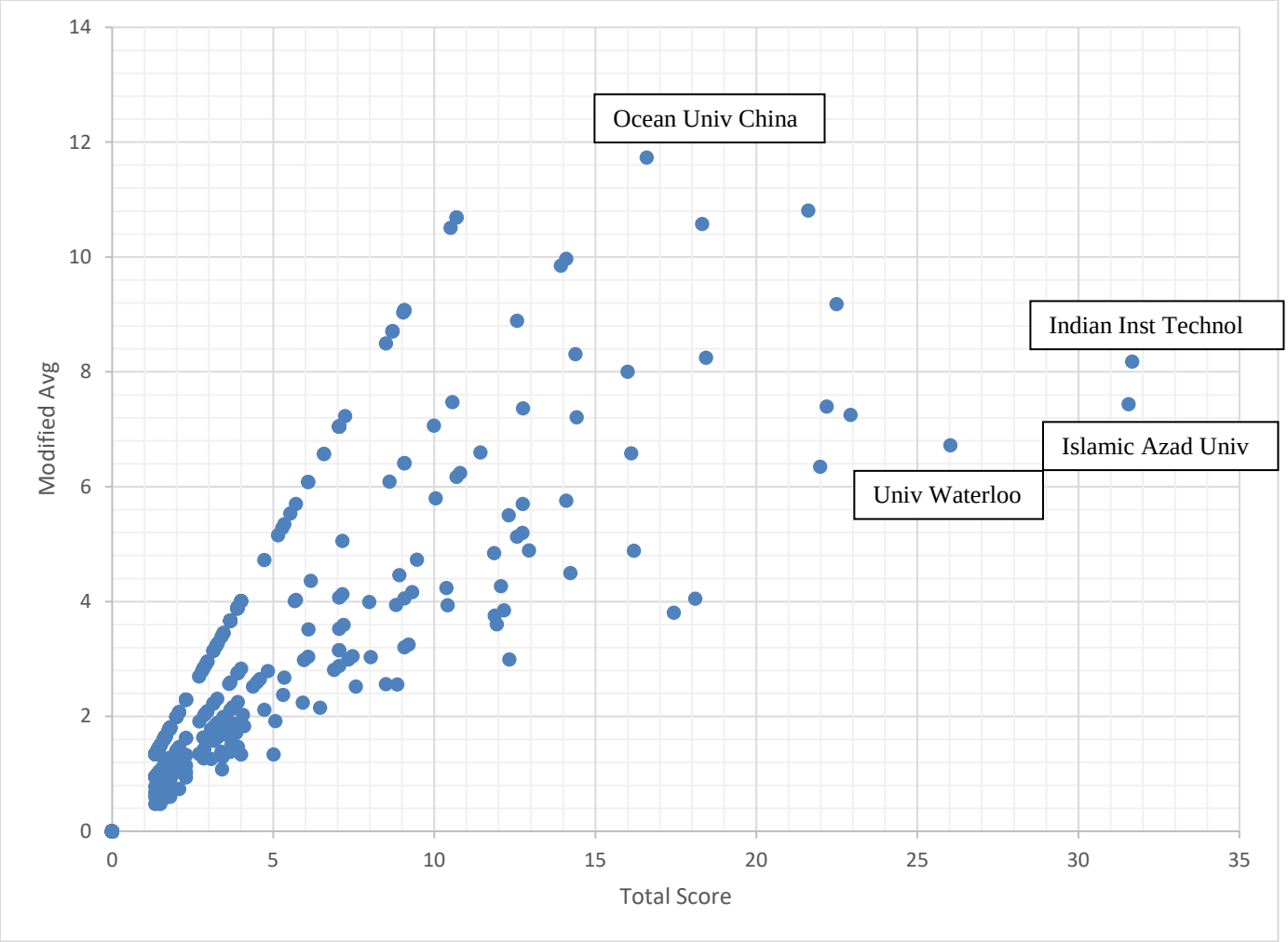

Figure 8-A. Normalized vs. Total EScores for Leading Organizations Publishing on Non-Linear Programming

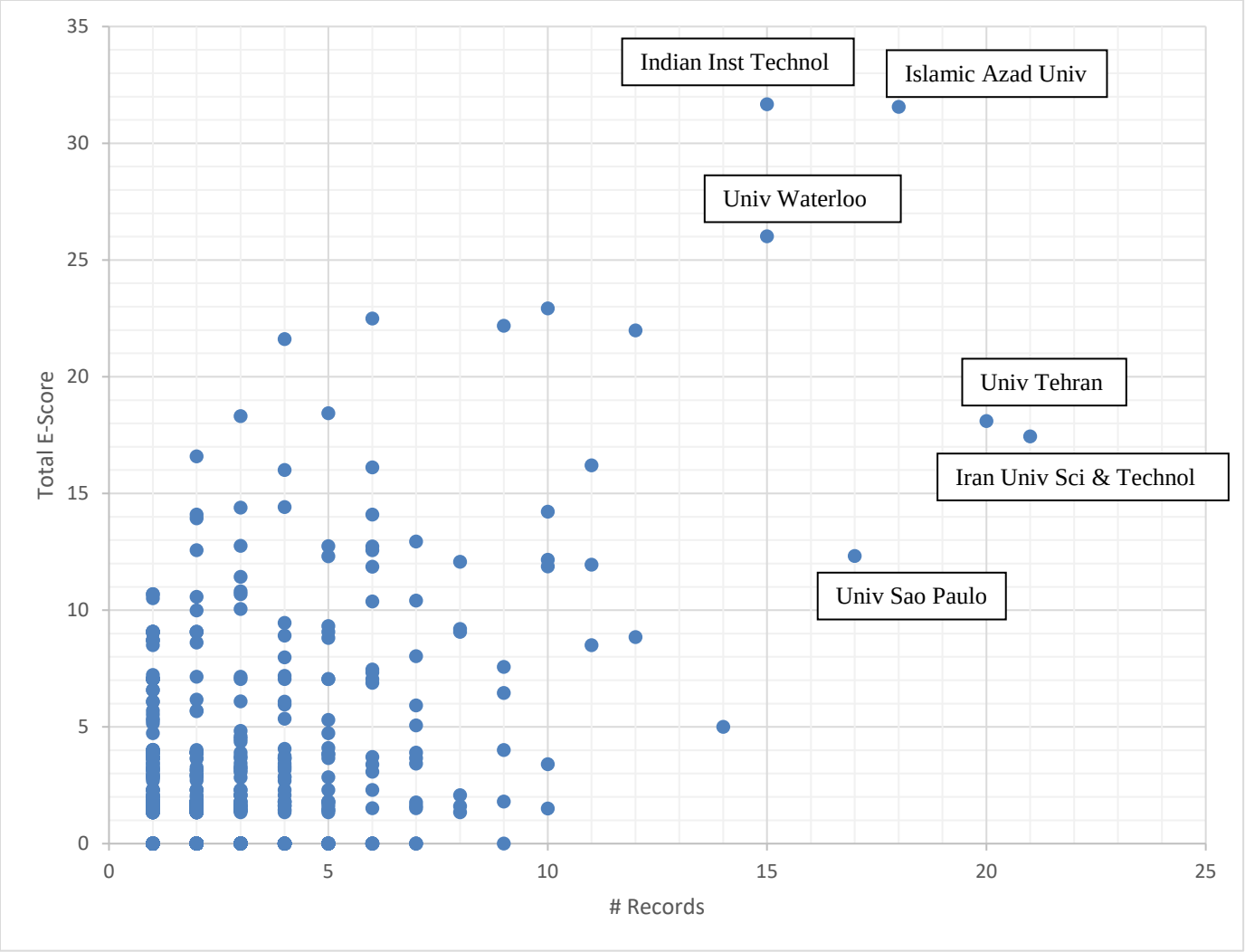

Figure 9-A. Total EScores for Leading Organizations Publishing on Non-Linear Programming by Number of Publications

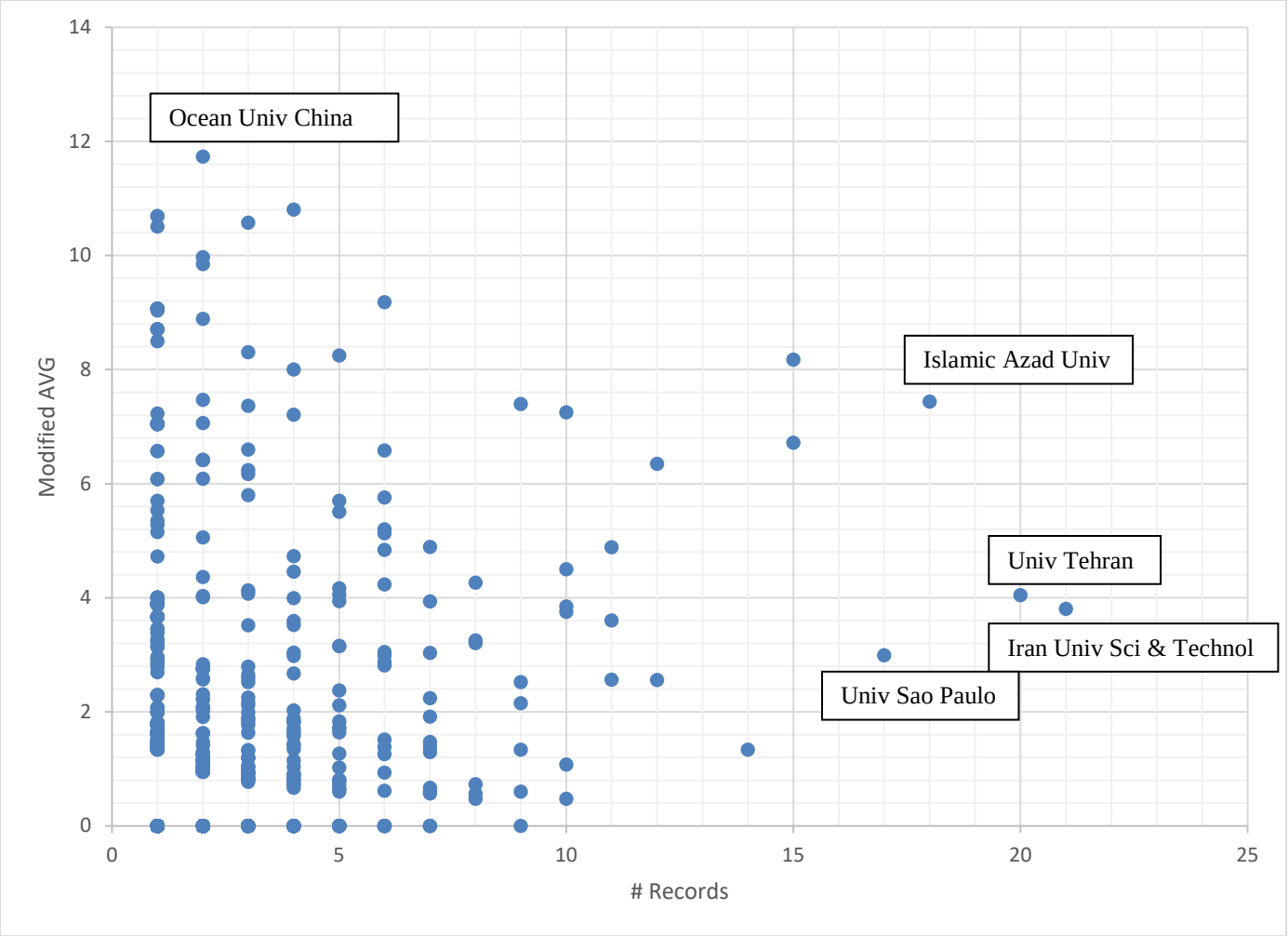

Figure 10-A. Normalized EScores for Leading Organizations Publishing on Non-Linear Programming by Number of Publications

Cutting-Edge Authors: Graphs for Total and Normalized EScoring

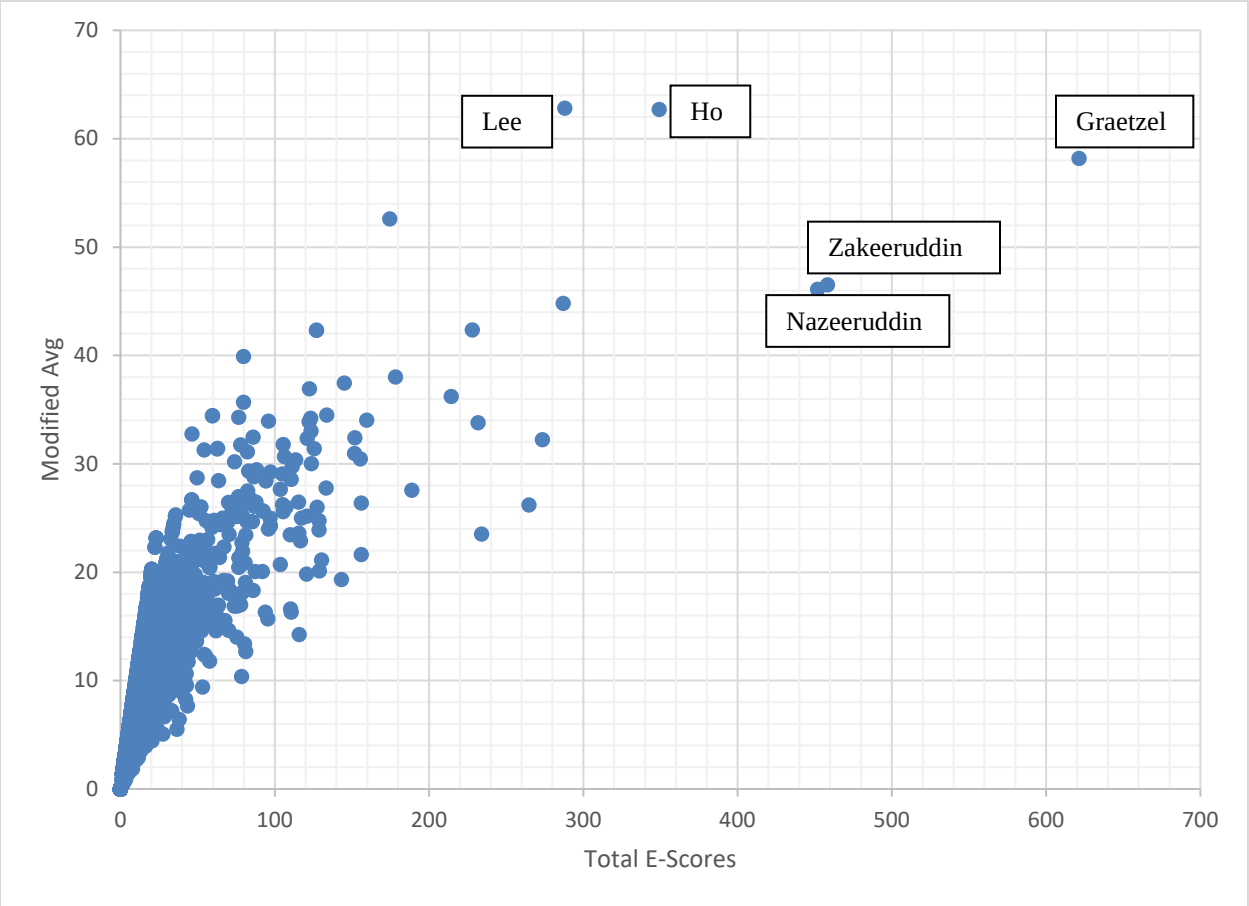

Figure 11-A. Normalized vs. Total EScores for Leading Authors Publishing on DSSCs

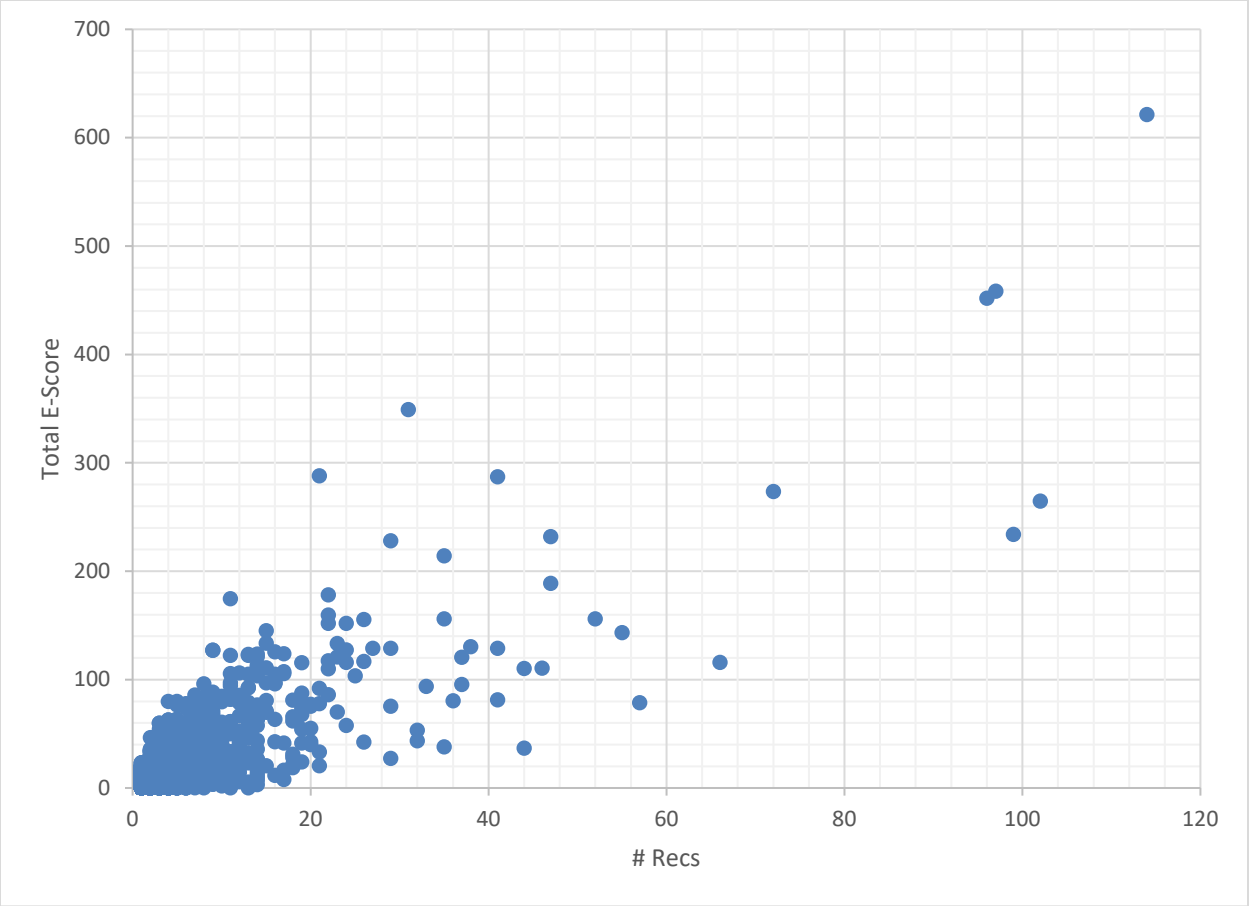

Figure 12-A. Total EScores for Leading Authors Publishing on DSSCs by Number of Publications

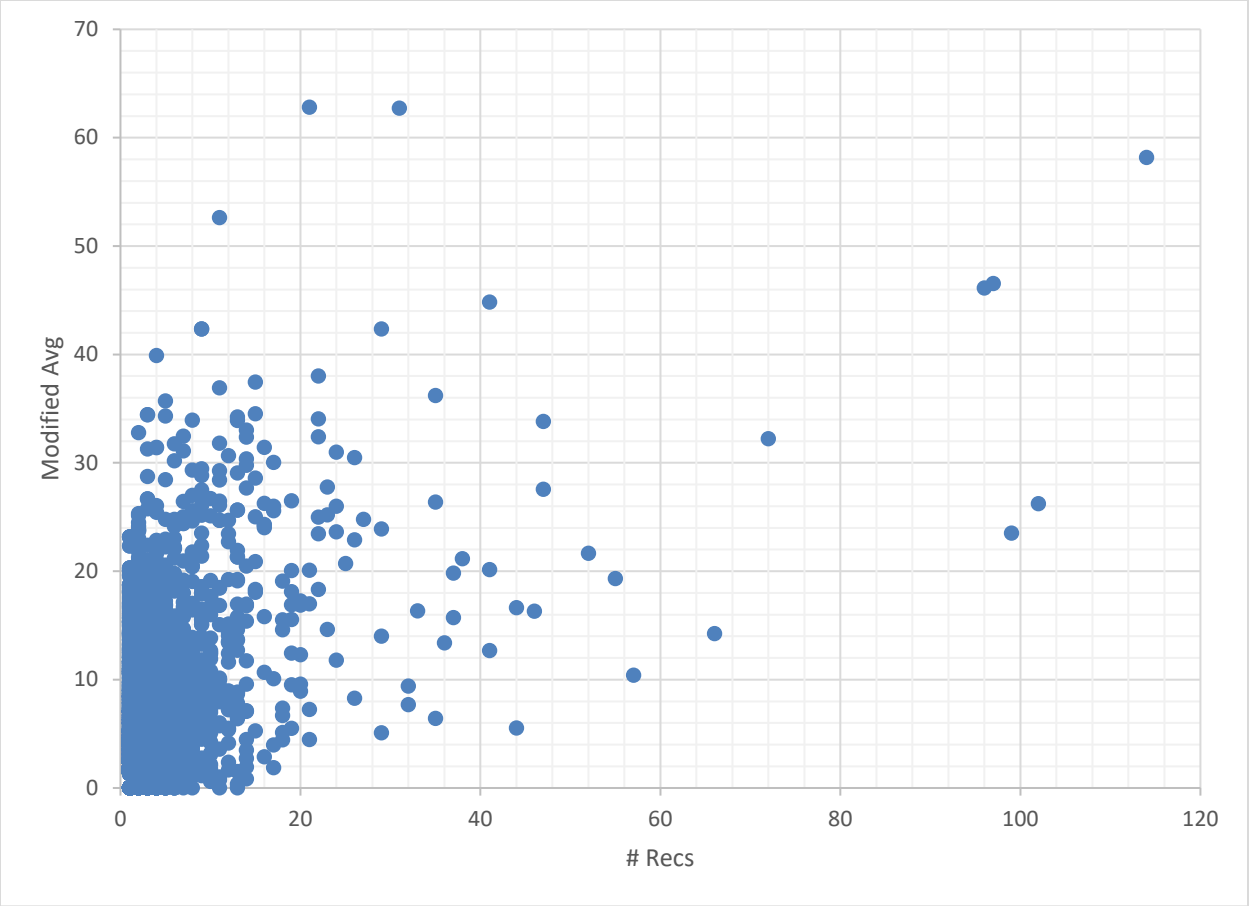

Figure 13-A. Normalized EScores for Leading Authors Publishing on DSSCs by Number of Publications

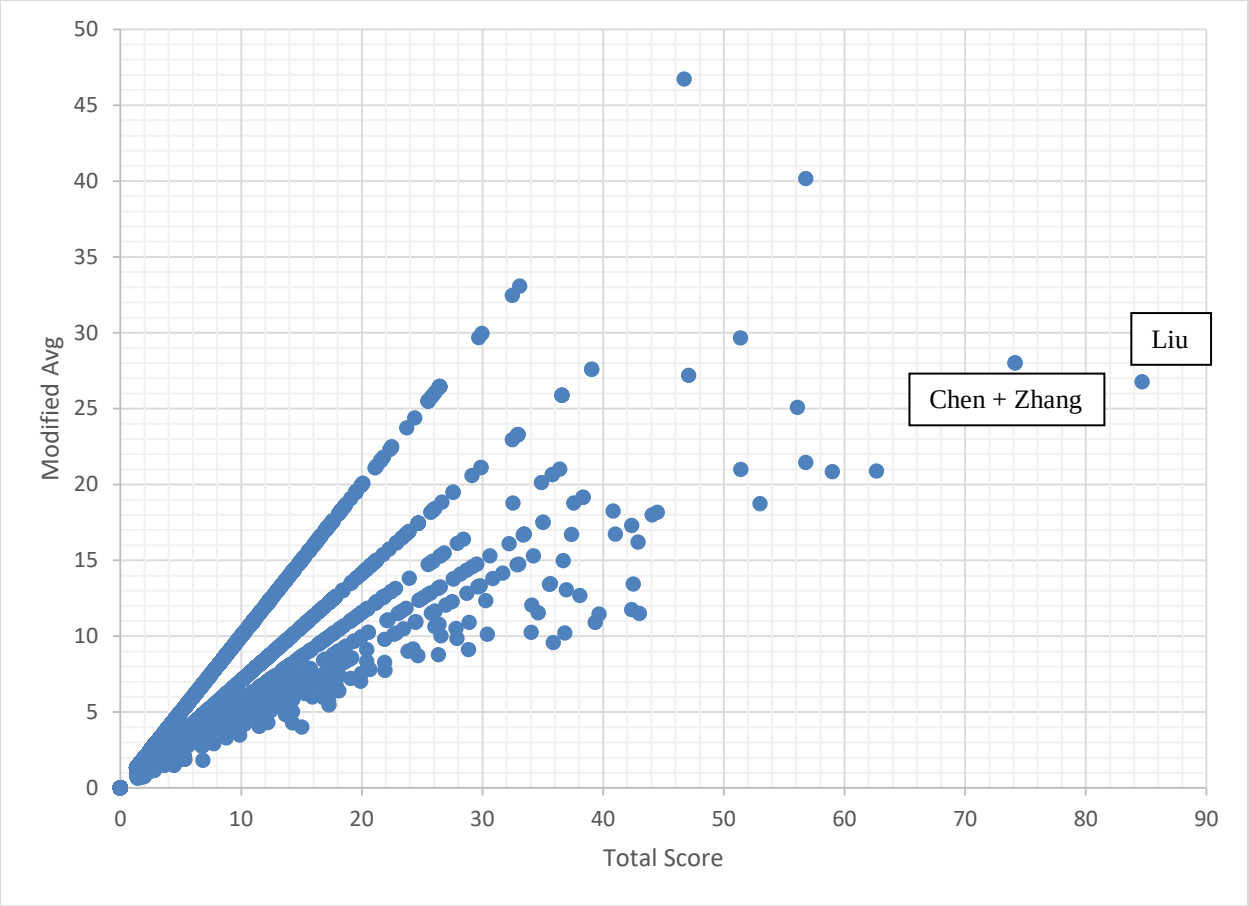

Figure 14-A. Normalized vs. Total EScores for Leading Authors Publishing on Big Data

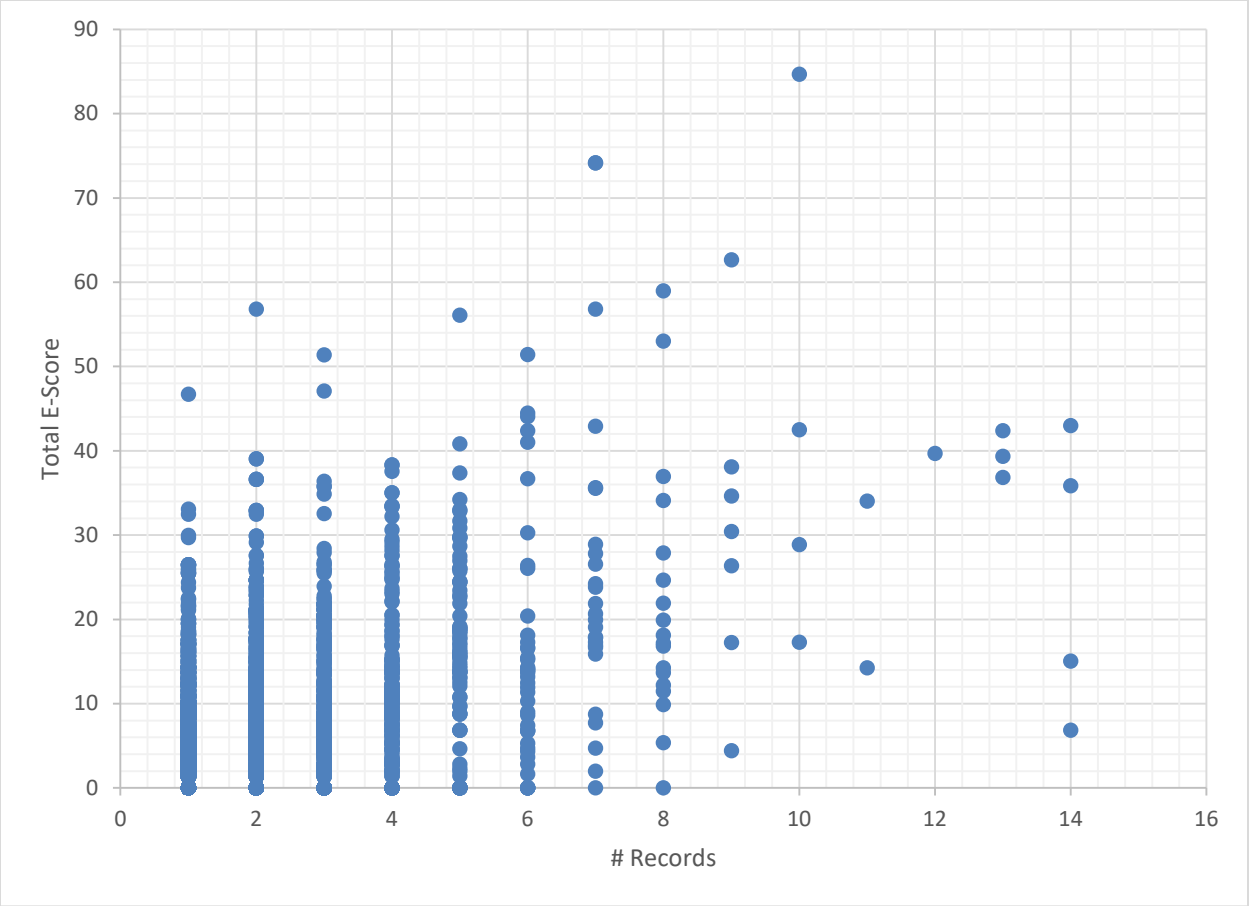

Figure 15-A. Total EScores for Leading Authors Publishing on Big Data by Number of Publications

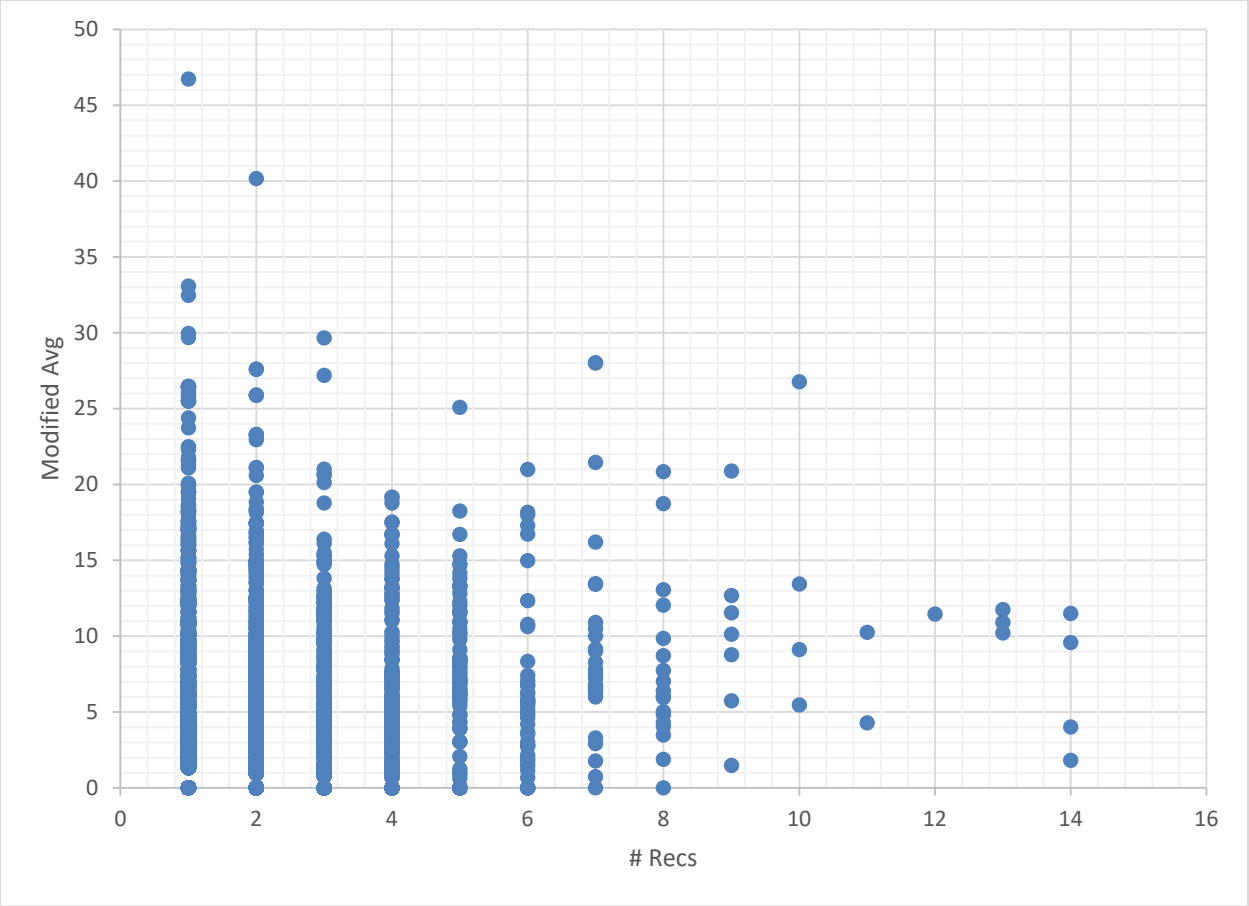

Figure 16-A. Normalized EScores for Leading Authors Publishing on Big Data by Number of Publications

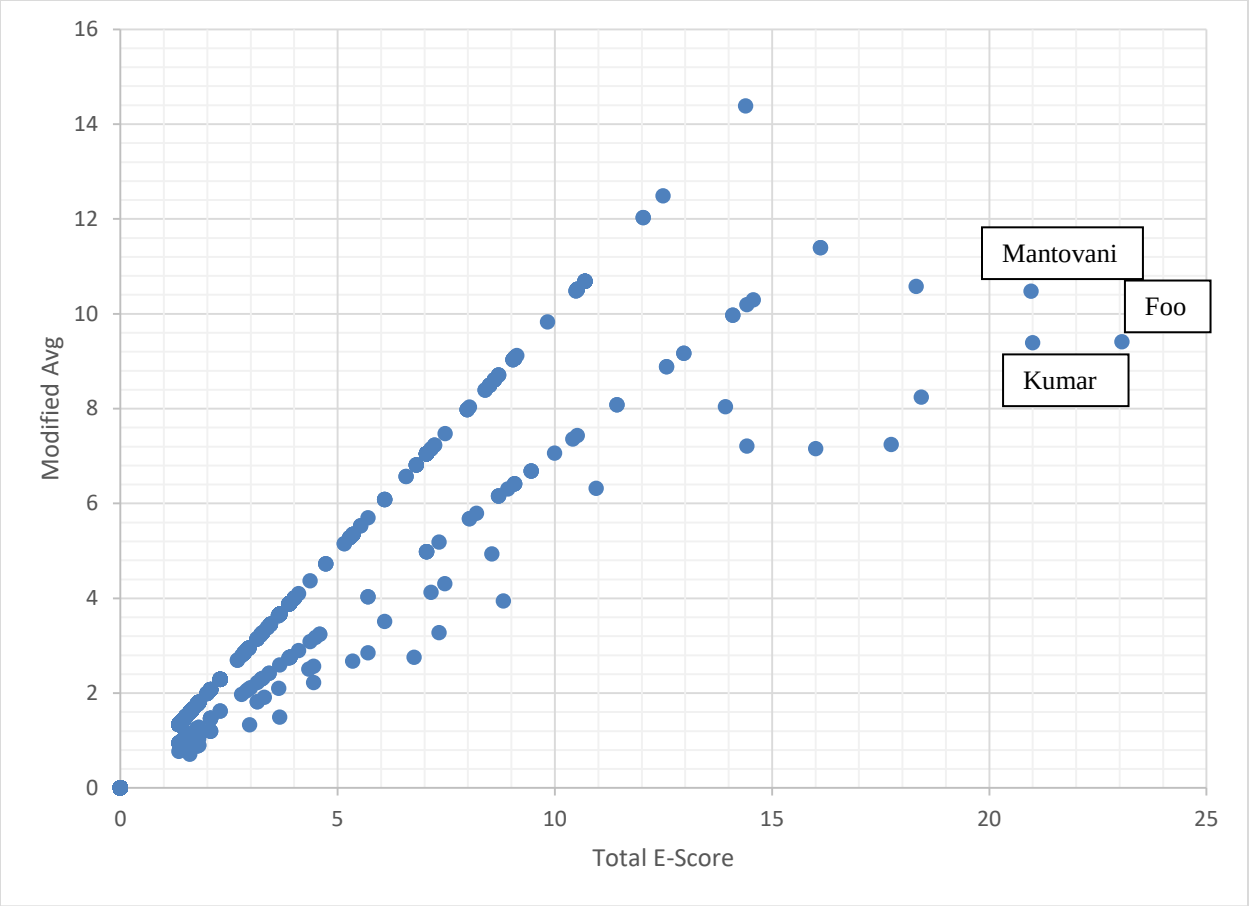

Figure 17-A. Normalized vs. Total EScores for Leading Authors Publishing on Non-Linear Programming

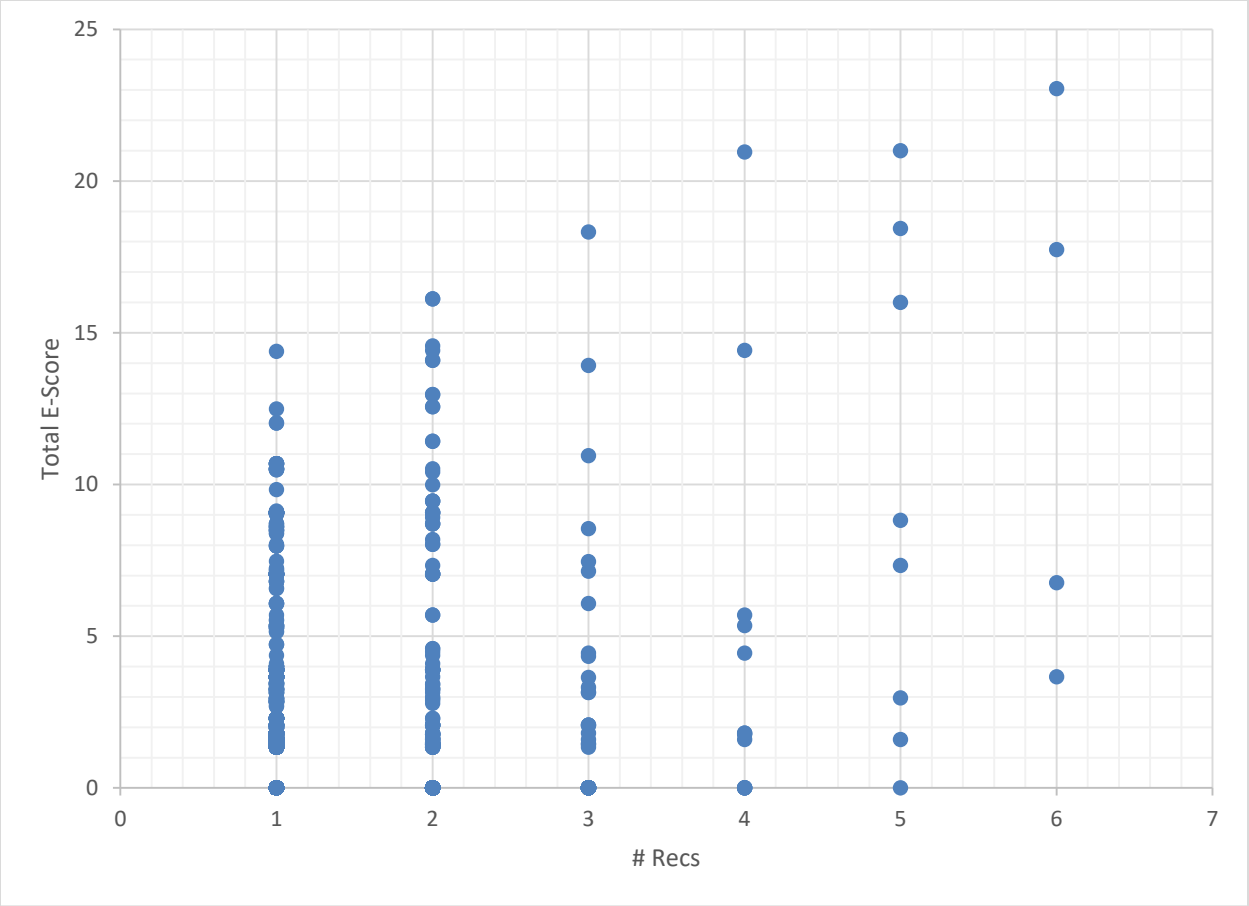

Figure 18-A. Total EScores for Leading Authors Publishing on Non-Linear Programming by Number of Publications

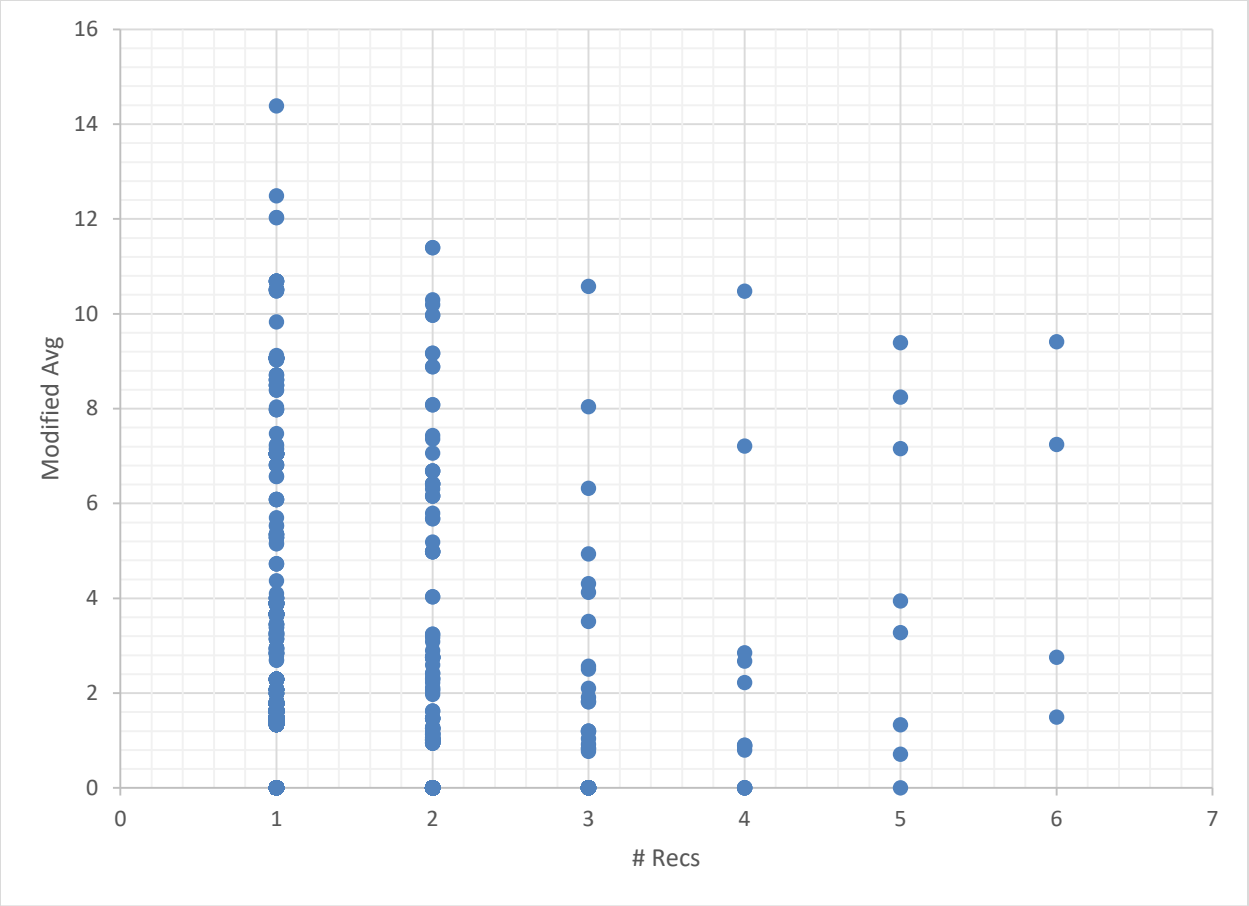

Figure 19-A. Normalized EScores for Leading Authors Publishing on Non-Linear Programming by Number of Publications

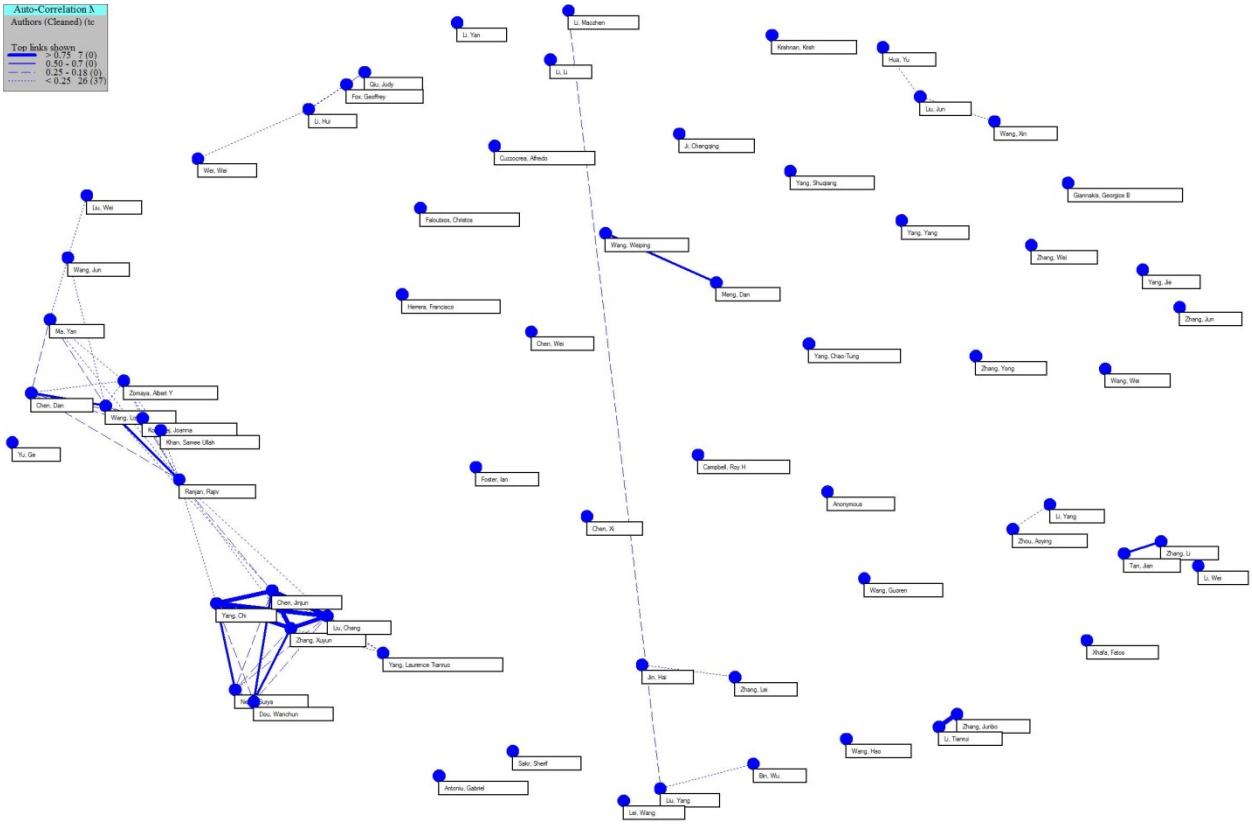

Figure 20-A. Collaboration among Prolific Big Data Authors

Visualizations of Country-level Total and Normalized EScoring for the 3 Test Datasets

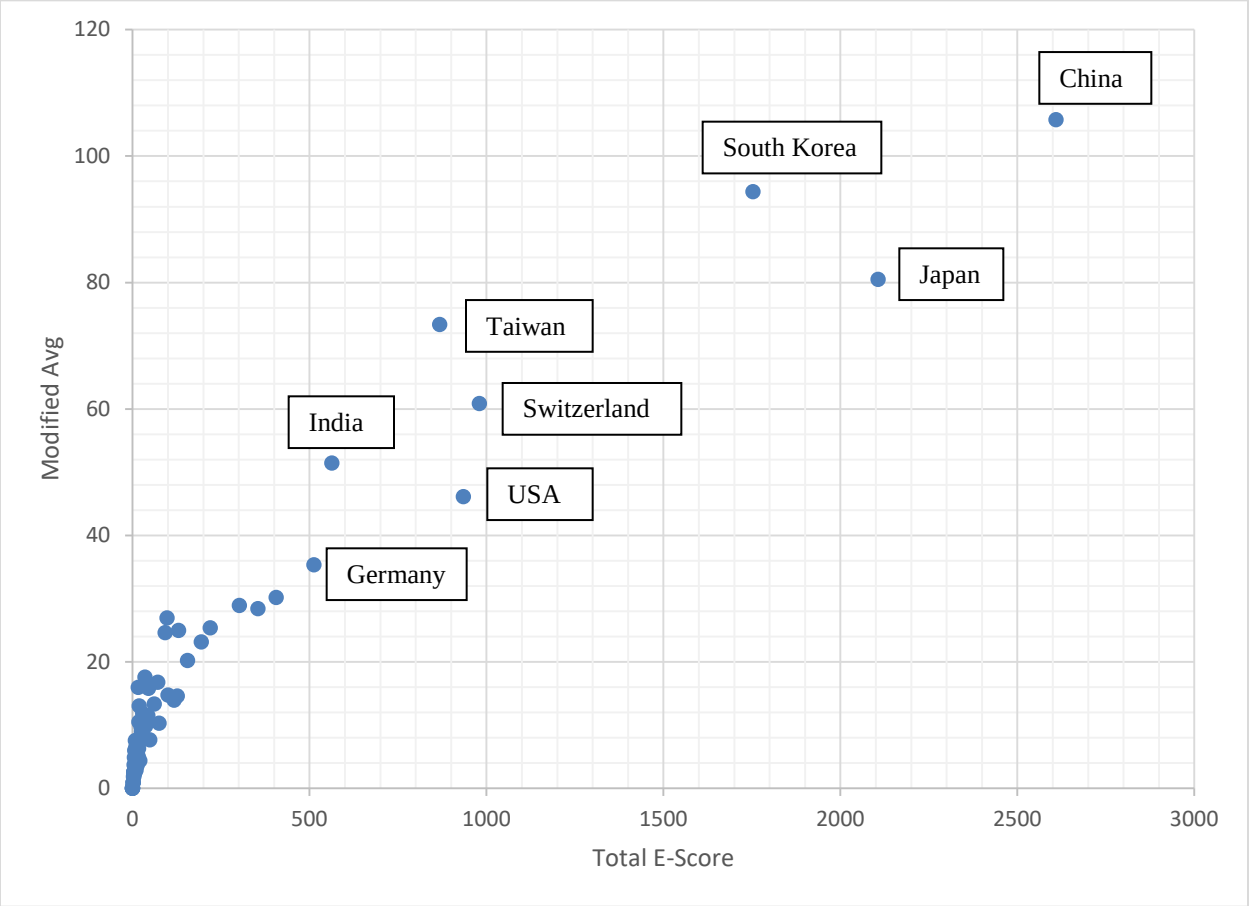

Figure 21-A. Normalized vs. Total EScores for Leading Countries Publishing on DSSCs

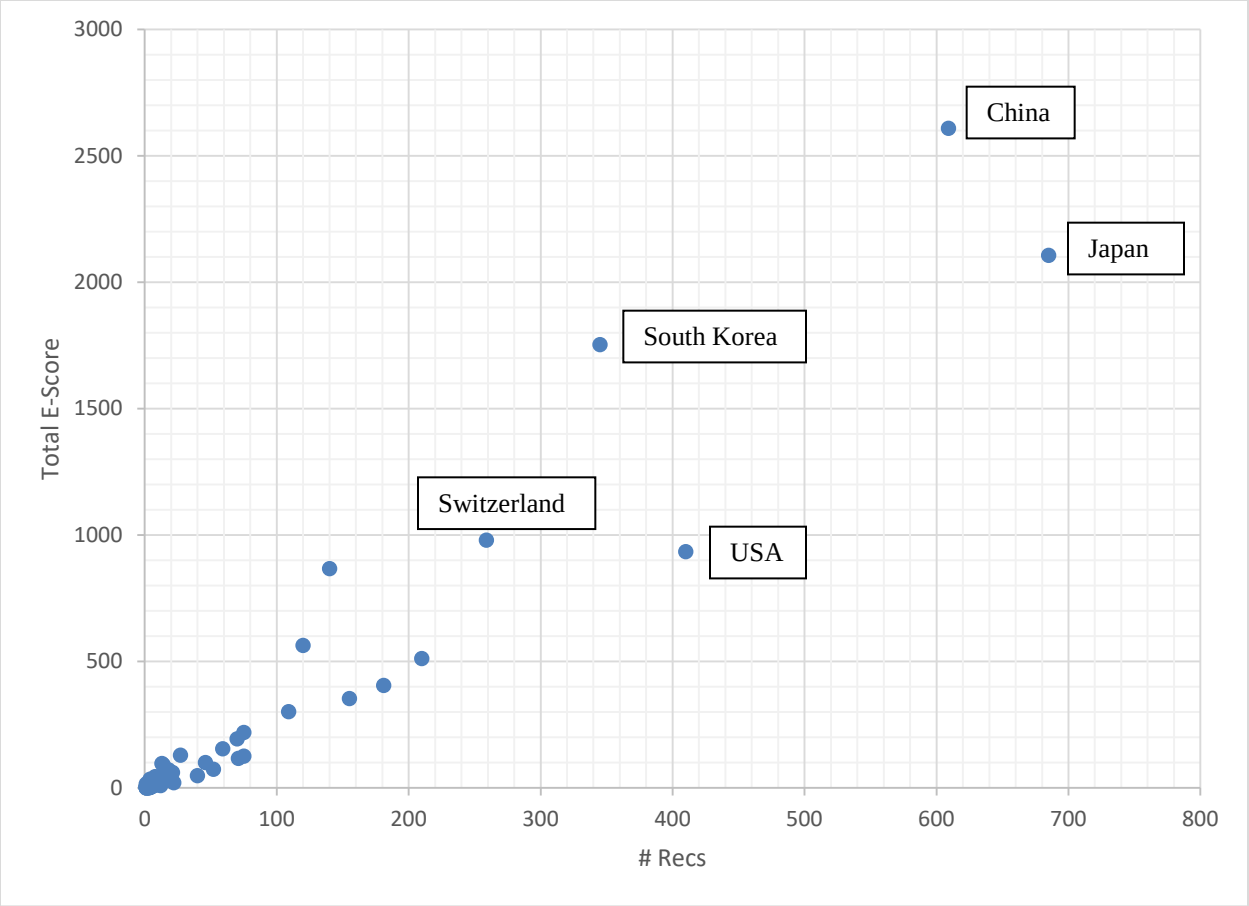

Figure 22-A. Total EScores for Leading Countries Publishing on DSSCs by Number of Publications

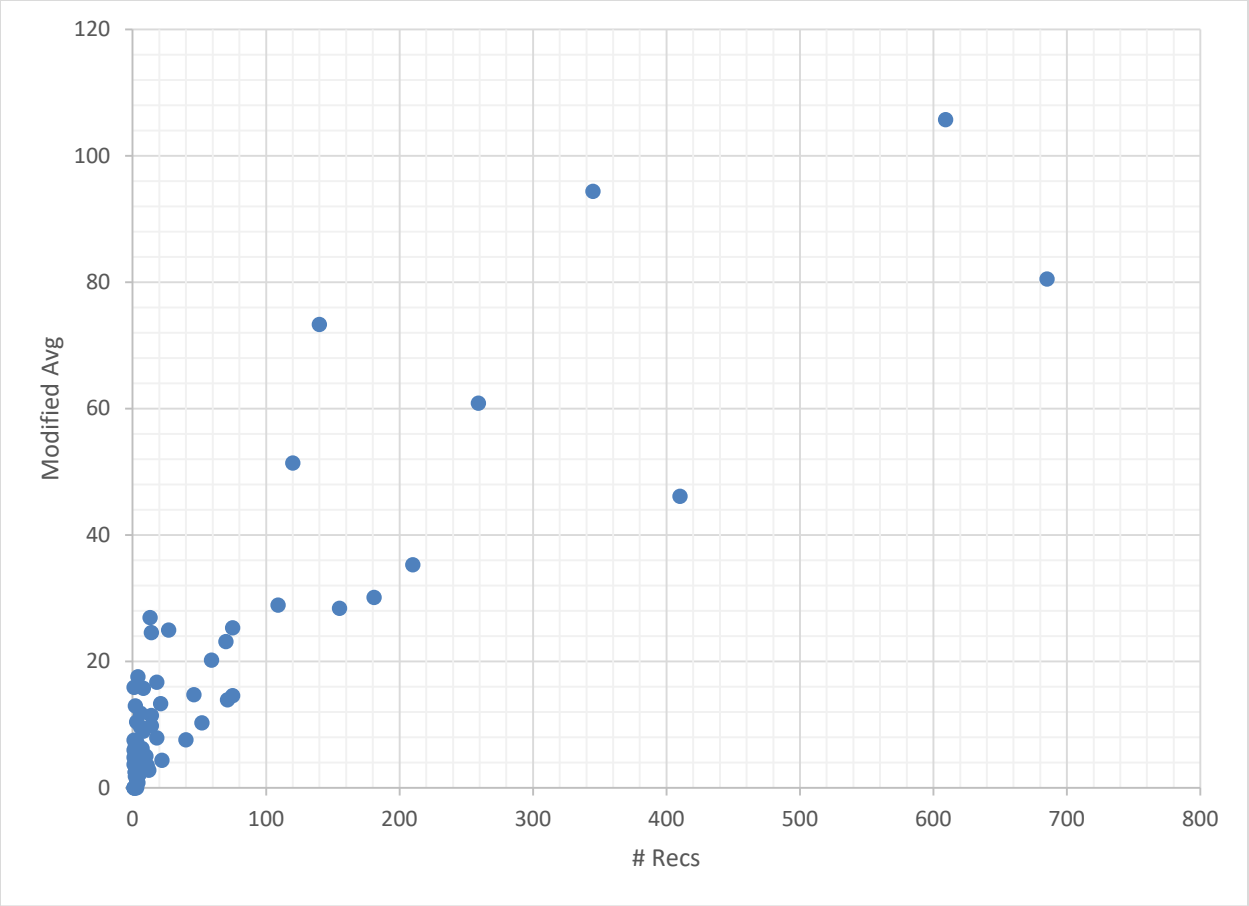

Figure 23-A. Normalized EScores for Leading Countries Publishing on DSSCs by Number of Publications

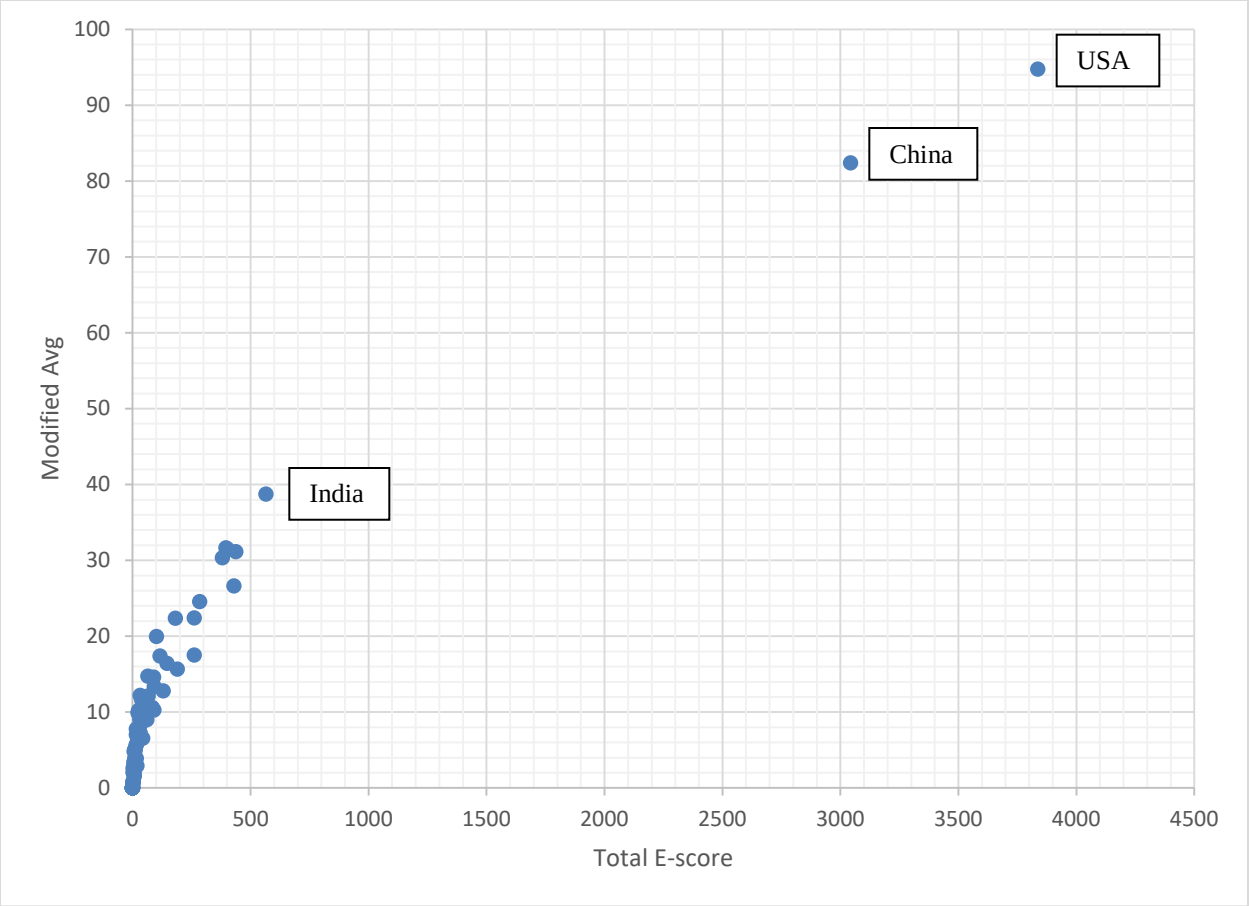

Figure 24-A. Normalized vs. Total EScores for Leading Countries Publishing on Big Data

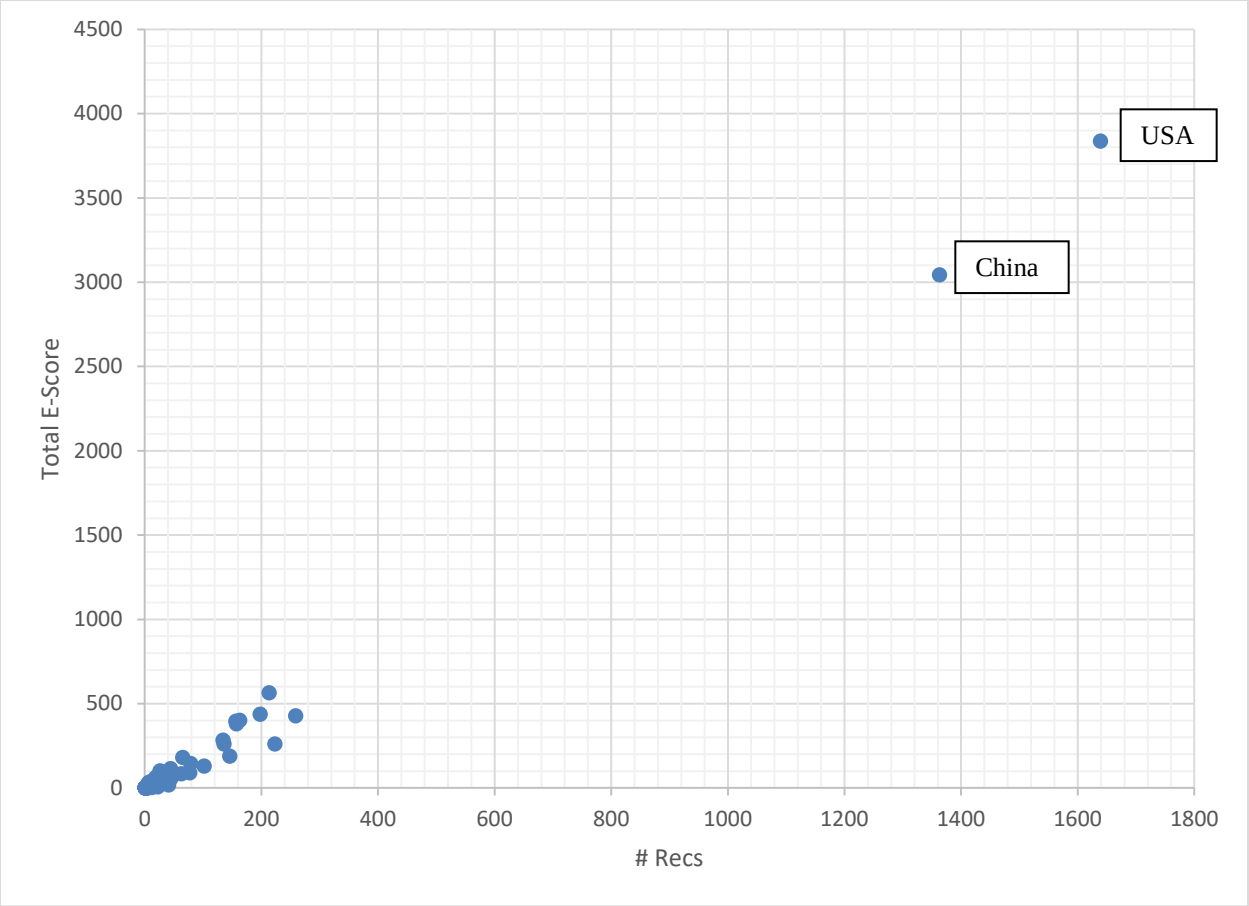

Figure 25-A. Total EScores for Leading Countries Publishing on Big Data by Number of Publications

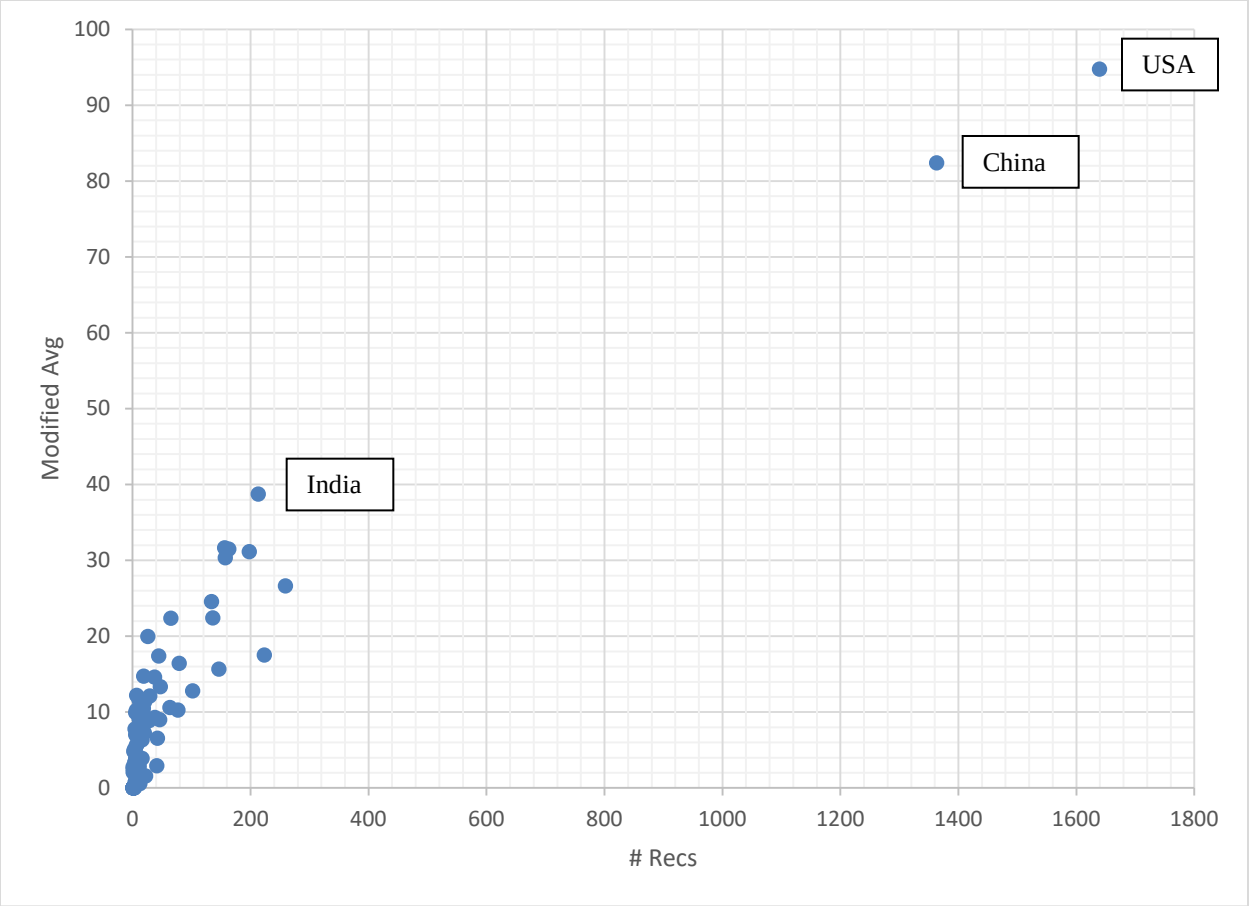

Figure 26-A. Normalized EScores for Leading Countries Publishing on Big Data by Number of Publications

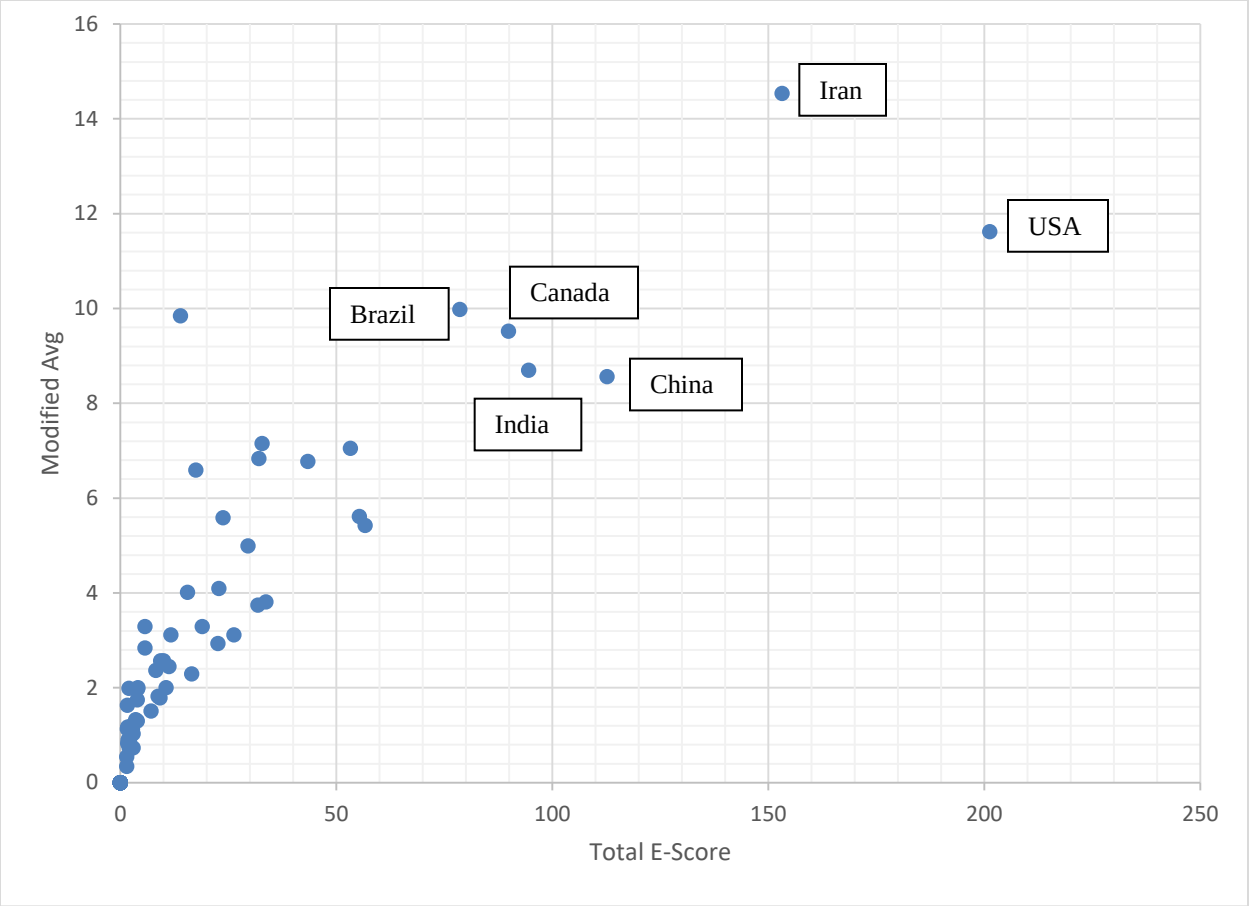

Figure 27-A. Normalized vs. Total EScores for Leading Countries Publishing on Non-Linear Programming

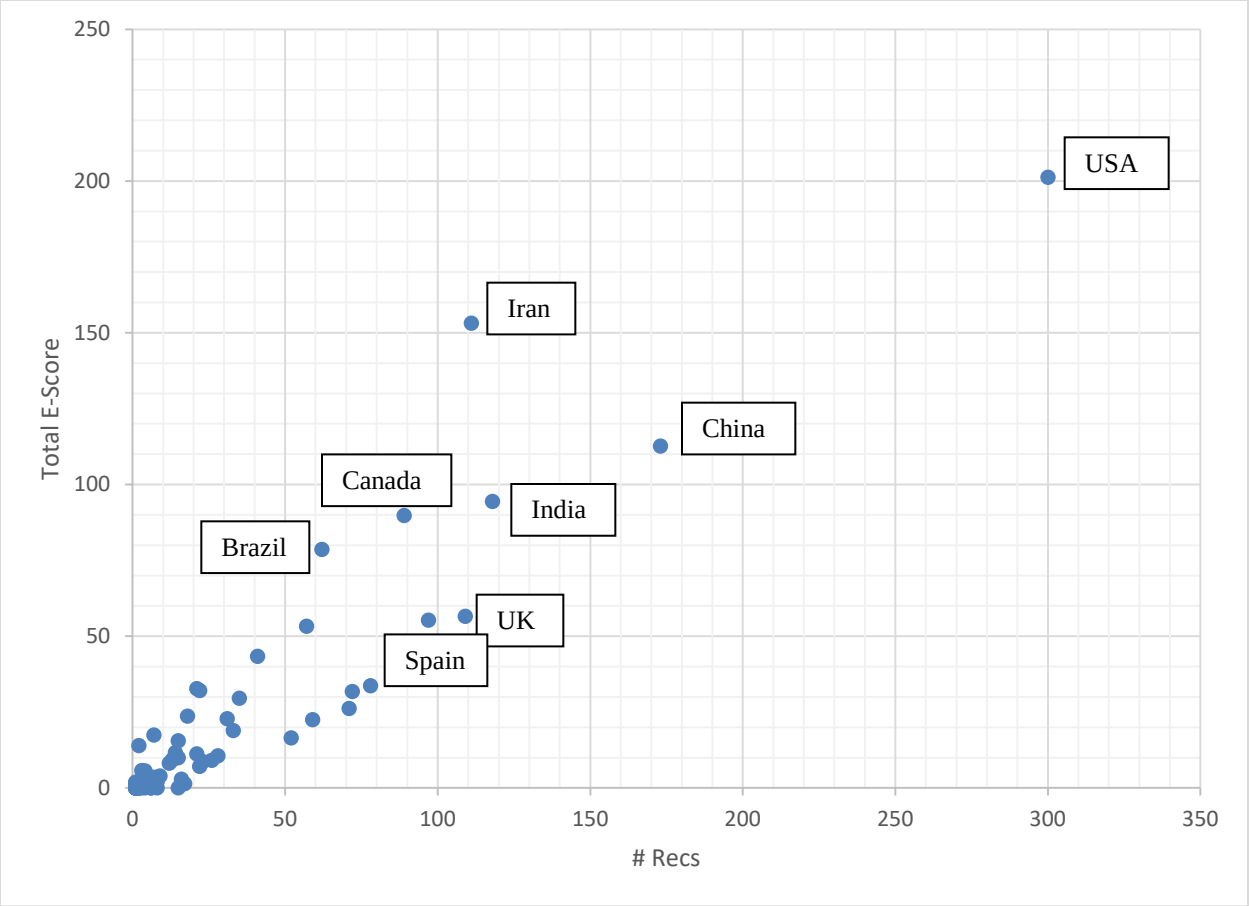

Figure 28-A. Total EScores for Leading Countries Publishing on Non-Linear Programming by Number of Publications

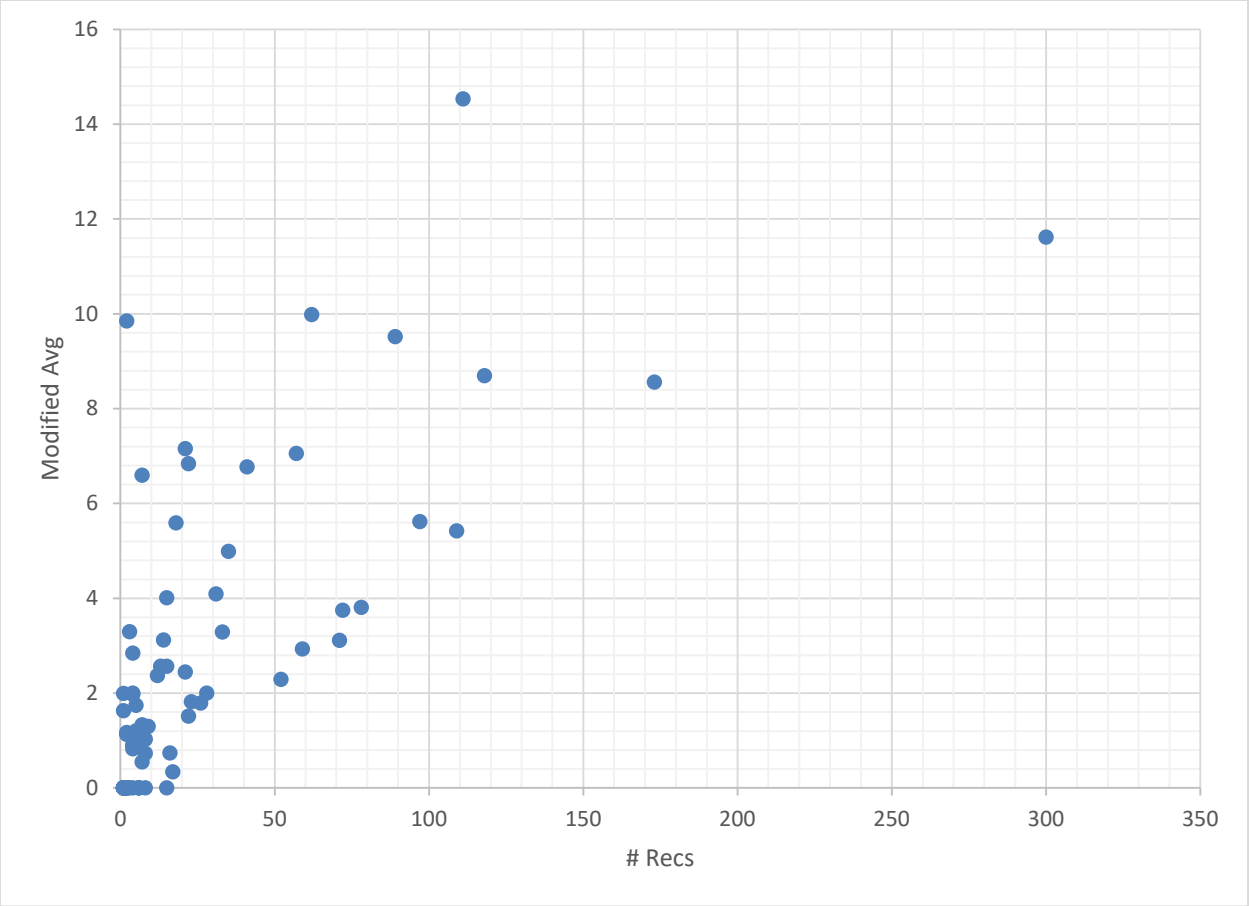

Figure 29-A. Normalized EScores for Leading Countries Publishing on Non-Linear Programming by Number of Publications

NEDD Cutting Edge Players: US Organizations

Second-order Emergence Indicators: Emergent Players
4] We analyze 50,745 articles on NEDD indexed by Web of Science, generating emergent terms, and then our second-order indicators of emergence for countries, organizations, and authors. Table 1 illustrates for author organizations (as a work in progress; we are working to translate the results into charts and visuals to communicate the key facets of emergence). This chart shows U.S. organizations most actively incorporating emergent NEDD terms in their author keywords or Keywords-Plus from the Web of Science records. Not shown, of the 60 organizations reaching our threshold to be identified as emergent, 24 are Chinese, followed by the 15 American ones shown. This indicator suggests that China is a leader, not a follower, in cutting-edge NEDD research.

Table 1 shows that some of those U.S. organizations have been pushing NEDD frontiers (i.e., their articles are rich in our identified emergent terms) for both time periods. Others show up for 2009 and drop below threshold for 2012. Vice-versa, some organizations not present in 2009 "emerge" in 2012.

Table 1-E. Emergent U.S. Research Organizations in Nano-Enabled Drug Delivery

Emergent Organizations	Frontier in 2009	Top Emergent Term 2009	Frontier in 2012	Top Emergent Term 2012
Harvard Univ	Yes	6 | Magnetic nanoparticle	Yes	7 | iron oxide nanoparticle
Univ Washington	Yes	10 | quantum dot	Yes	15 | iron oxide nanoparticle
MIT	Yes	9 | quantum dot	Yes	10 | cellular uptake
Univ Calif Los Angeles	Yes	6 | Mesoporous Silica	Yes	6 | Mesoporous Silica nanoparticle
Northwestern Univ	Yes	9 | Carbon Nanotube	Yes	9 | cellular uptake
Emory Univ	Yes	7 | quantum dot	Yes	9 | iron oxide nanoparticle
Univ Michigan	Yes	6 | quantum dot	No	N/A
Stanford Univ	Yes	6 | quantum dot	No	N/A
Univ Massachusetts	Yes	9 | Gold Nanoparticle	No	N/A
Georgia Inst Technol	Yes	7 | quantum dot	No	N/A
Northeastern Univ	Yes	7 | Nanocarrier	No	N/A
Univ Utah	No	N/A	Yes	6 | cellular uptake; pH; siRNA delivery
Univ Calif San Diego	No	N/A	Yes	13 | Mesoporous Silica nanoparticle
Purdue Univ	No	N/A	Yes	7 | siRNA delivery
Rutgers State Univ	No	N/A	Yes	7 | cellular uptake

Notes: Numbers in the "Top Emergent Term" columns indicate how many articles contained that most frequent term for the organization in that time period. This provides a simple pointer to the prevalent research thrusts.