id	Name	Affiliation	Areas of Interest	#Articles	#Citations
G1	Robert Smith	University of California	Artificial Intelligence, Text Mining	25	1602
G2	Joan Clarke	University of Buenos Aires	Entomology	12	441
G3	Anthony H. Kane	City, University of London	Database	9	41
G4	Joe Green	PSL University, Paris	Computer Science, Algorithms	149	6221
G5	Joanne Clark	University of Buenos Aires	Entomology	12	429
G6	Annabelles Greenwood	University of Toronto	Algorithms	2	1
G7	Robert Smith	University of California	Database, Text Mining	26	1610
G8	Antony Kane	Unknown	Biological Databases	9	39
G9	Serge Lenglet	New York University	Entomology	22	2291
G10	Antony Kane	City, University of London	Bioinformatics	5	26

(a)

id	Key
G1	thArt1
G2	keEnt4
G3	neDat4
G4	enCom6
G5	rkEnt4
G6	odAlg1
G7	thDat1
G8	neBio3
G9	etEnt2
G10	neBio2

(b)

id	Key
G1	thAr, hArt, Art1
G2	keEn, eEnt, Ent4
G3	neDa, eDat, Dat4
G4	enCo, nCom, Com6
G5	rkEn, kEnt, Ent4
G6	odAl, dAlg, Alg1
G7	thDa, hDat, Dat1
G8	neBi, eBio, Bio3
G9	etEn, tEnt, Ent2
G10	neBi, eBio, Bio2

(c)

Ent4
G2
G5

neBi
G8
G10

eBio
G8
G10

(d)

Figure 3.3: Applying Standard and 4-grams Blocking to the Dirty DS of Figure 2.2: (a) the input DS with highlighted the information used in blocking keys, (b) the blocking keys of Standard Blocking per profile, (c) the blocking keys of 4-grams Blocking per profile, and (d) the blocks of 4-grams Blocking—Standard Blocking yields no blocks.