Noname manuscript No.
(will be inserted by the editor)

Supplementary material on the paper "Schema Evolution and Foreign Keys: a Study on Usage, Heartbeat of Change and Relationship of Foreign Keys to Table Activity"

Panos Vassiliadis · Michail-Romanos Kolozoff · Maria Zerva · Apostolos V. Zarras

the date of receipt and acceptance should be inserted later

1 Supplements on how tables and foreign keys evolve

1.1 Different metrics for the evolution of tables and foreign keys

One of the first problems that we had to address in our study is the multitude of the measures involved in the study of the evolution of foreign keys.

Vertex-related measures. The fundamentals vertex-related measures that we have used are (a) the family of *vertex degrees* (*in-*, *out-* and *total degree*), and (b) simple vertex-related measures like the *clustering coefficient* and the *betweenness centrality* of a vertex. For each table, for each version of the schema history, we have measured the respective values of these measures. Moreover, we have computed the respective values for the Diachronic Graph.

Variants of measures. Then, the problem that arises involves selecting representative values for each table and for each measure. We have resorted to 6 representative values to characterize each measure, for each table, and specifically: (a) *measureDG,* (b) *measureBirth*, (c) *measureEnd*, (d) *measureAvg,* (e) *measureMax,* (f) *measureMin, w*here the suffix of the measure's name, signifies the value of the measure for the respective node as measured (a) in the Diachronic Graph, (b) in the first version where the table appeared, (c) in the last known version of the table, i.e., the final version of the database history for survivors and the time of death for deleted tables, (d) the average value for the versions where the table was part of the schema, (e) the maximum value ever attained by the table in its

P. Vassiliadis, A. Zarras
Dept. of Computer Science & Engineering, Univ. of Ioannina, Greece
E-mail: {pvassil, zarras}@cs.uoi.gr

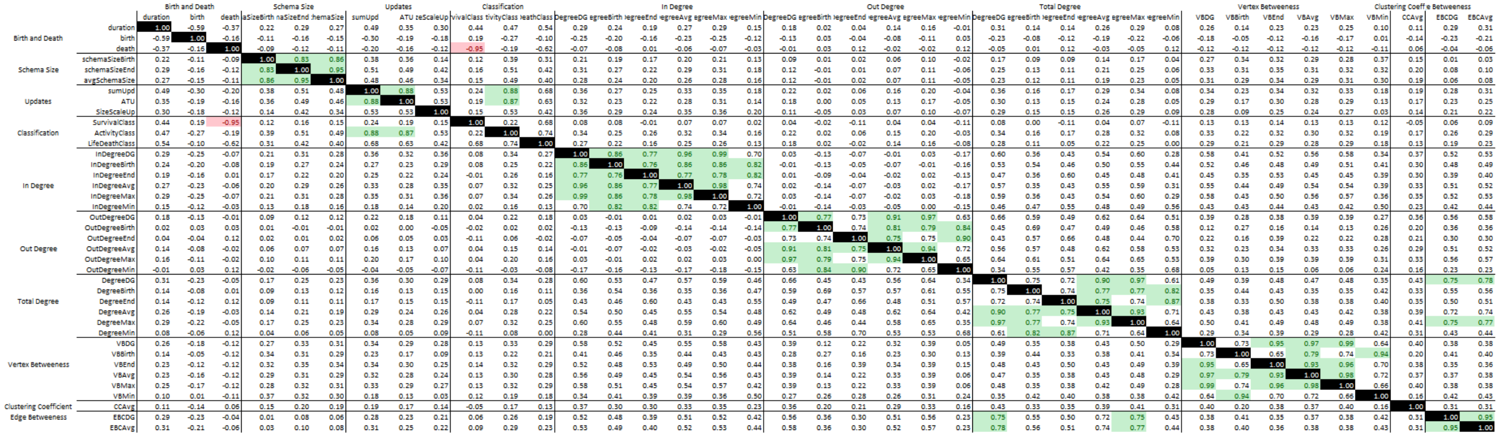
M. Kolozoff
Advantage FSE, Athens, Greece E-mail: mkolozoff@gmail.com
Work done while in the Univ. of Ioannina

M. Zerva
P&I AG, Ioannina, Greece E-mail: mariazerva92@gmail.com
Work done while in the Univ. of Ioannina

entire life, (f) the minimum such value, respectively. All these alternatives for each measures are collectively refered to as *variants* of the measure.

Apart from all these vertex-related measures, we have also measured, for each table, the *maximum* value of the *edge betweenness centrality* of *all* its inciting edges. Thus, for each version, each node comes with its maximum edge betweenness centrality (practically, converting an edge-related measure to a vertex-related one). Then, we averaged this value over all the history of the schema, producing the *average edge betweenness centrality* (EBCAvg) for a table. We also measured each table's *edge betweenness centrality at the Diachronic Graph* (EBCDG).

Table properties related to evolution. Apart from the properties of the tables that are related to foreign keys or their graph representation, there are also properties that relate to their birth, possible death and activity in-between. For each table, we have extracted several families of properties (see [1, 3] for very detailed definitions). Concerning *birth and death*, we have extracted the version of *birth* and (if applicable) *death*, the *Last Known Version* (LKV) of the table, and its *duration* (in number of versions present in the schema history). Concerning the *schema size* of a table (in attributes), we have extracted the *schema size at birth*, *at LKV*, and *on average*. We also measure size *scaleUp* as the fraction of the schema at LKV over the schema size at birth. Concerning the amount of update activity we have counted the *sum of updates* to a table (including attribute ejections, injections, data type and primary key updates – all measured by the number of involved attributes) and the *count of versions holding table updates* to the table. We define the *Average Transitional Update rate* (ATU) as the fraction of sum of updates over duration (in number of versions) and the *Update Rate* as the fraction of count of versions with updates over duration.



		Birth and Death			Schema Size			Updates			Classification			In Degree						Out Degree						Total Degree						Vertex Betweeness						Clustering Coeff e Betweeness		
		duration	birth	death	aSizeBirth	naSizeEnd	hemaSize	sumUpd	ATU	zeScaleUp	vivalClass	tivityClass	eathClass	DegreeDG	egreeBirth	egreeEnd	egreeAvg	egreeMax	egreeMin	DegreeDG	egreeBirth	egreeEnd	egreeAvg	egreeMax	egreeMin	DegreeDG	egreeBirth	egreeEnd	egreeAvg	egreeMax	egreeMin	VBDG	VBBirth	VBEnd	VBAvg	VBMax	VBMin	CCAvg	EBCDG	EBCAvg
	duration	1.00	-0.59	-0.37	0.22	0.29	0.27	0.49	0.35	0.30	0.44	0.47	0.54	0.29	0.24	0.19	0.27	0.29	0.15	0.18	0.02	0.04	0.14	0.16	-0.01	0.31	0.14	0.14	0.26	0.29	0.08	0.26	0.14	0.23	0.23	0.25	0.10	0.11	0.29	0.31
Birth and Death	birth	-0.59	1.00	-0.16	-0.11	-0.16	-0.15	-0.30	-0.19	-0.18	0.19	-0.27	-0.10	-0.25	-0.20	-0.16	-0.23	-0.25	-0.12	-0.13	0.03	-0.04	-0.08	-0.11	0.03	-0.23	-0.08	-0.12	-0.19	-0.22	-0.06	-0.18	-0.05	-0.12	-0.16	-0.17	0.01	-0.14	-0.23	-0.21
	death	-0.37	-0.16	1.00	-0.09	-0.12	-0.11	-0.20	-0.16	-0.12	-0.95	-0.19	-0.62	-0.07	-0.08	0.01	-0.06	-0.07	-0.03	-0.01	0.03	0.12	-0.02	-0.02	0.12	-0.05	0.01	0.12	-0.03	-0.05	0.12	-0.12	-0.12	-0.12	-0.12	-0.12	-0.11	0.06	-0.04	-0.06
	schemaSizeBirth	0.22	-0.11	-0.09	1.00	0.83	0.86	0.38	0.36	0.14	0.12	0.39	0.31	0.21	0.19	0.17	0.20	0.21	0.13	0.09	0.01	0.02	0.06	0.10	-0.02	0.17	0.09	0.09	0.14	0.17	0.04	0.27	0.34	0.32	0.29	0.28	0.37	0.15	0.01	0.03
Schema Size	schemaSizeEnd	0.29	-0.16	-0.12	0.83	1.00	0.95	0.51	0.49	0.42	0.16	0.51	0.42	0.31	0.27	0.22	0.29	0.31	0.18	0.12	-0.01	0.01	0.07	0.11	-0.06	0.25	0.13	0.11	0.21	0.25	0.06	0.33	0.31	0.35	0.31	0.32	0.32	0.20	0.08	0.10
	avgSchemaSize	0.27	-0.15	-0.11	0.86	0.95	1.00	0.48	0.46	0.34	0.15	0.49	0.40	0.28	0.24	0.20	0.26	0.28	0.16	0.12	-0.01	0.02	0.07	0.11	-0.05	0.23	0.12	0.11	0.19	0.23	0.05	0.31	0.29	0.34	0.29	0.31	0.30	0.19	0.06	0.08
	sumUpd	0.49	-0.30	-0.20	0.38	0.51	0.48	1.00	0.88	0.53	0.24	0.88	0.68	0.36	0.27	0.25	0.33	0.35	0.18	0.22	0.02	0.06	0.16	0.20	-0.04	0.36	0.16	0.17	0.29	0.34	0.08	0.34	0.23	0.34	0.32	0.33	0.18	0.19	0.28	0.31
Updates	ATU	0.35	-0.19	-0.16	0.36	0.49	0.46	0.88	1.00	0.53	0.19	0.87	0.63	0.32	0.23	0.22	0.28	0.31	0.14	0.18	0.00	0.05	0.13	0.17	-0.05	0.30	0.13	0.15	0.24	0.28	0.05	0.29	0.17	0.30	0.28	0.29	0.13	0.17	0.23	0.25
	SizeScaleUp	0.30	-0.18	-0.12	0.14	0.42	0.34	0.53	0.53	1.00	0.15	0.53	0.42	0.36	0.29	0.24	0.35	0.36	0.20	0.11	-0.05	0.03	0.07	0.10	-0.07	0.29	0.15	0.15	0.26	0.29	0.09	0.28	0.09	0.25	0.24	0.27	0.03	0.14	0.21	0.22
	SurvivalClass	0.44	0.19	-0.95	0.12	0.16	0.15	0.24	0.19	0.15	1.00	0.22	0.68	0.08	0.08	-0.01	0.07	0.07	0.02	0.04	-0.02	-0.11	0.04	0.04	-0.11	0.08	0.00	-0.11	0.04	0.07	-0.11	0.13	0.13	0.14	0.13	0.13	0.12	-0.05	0.06	0.09
Classification	ActivityClass	0.47	-0.27	-0.19	0.39	0.51	0.49	0.88	0.87	0.53	0.22	1.00	0.74	0.34	0.25	0.26	0.32	0.34	0.16	0.22	0.02	0.06	0.15	0.20	-0.03	0.34	0.16	0.17	0.28	0.32	0.08	0.33	0.22	0.32	0.30	0.32	0.19	0.17	0.26	0.29
	LifeDeathClass	0.54	-0.10	-0.62	0.31	0.42	0.40	0.68	0.63	0.42	0.68	0.74	1.00	0.27	0.22	0.16	0.25	0.26	0.13	0.18	0.02	-0.02	0.14	0.16	-0.08	0.28	0.11	0.05	0.22	0.25	0.00	0.29	0.21	0.29	0.28	0.29	0.18	0.13	0.19	0.23
	InDegreeDG	0.29	-0.25	-0.07	0.21	0.31	0.28	0.36	0.32	0.36	0.08	0.34	0.27	1.00	0.86	0.77	0.96	0.99	0.70	0.03	-0.13	-0.07	-0.01	0.03	-0.17	0.60	0.36	0.43	0.54	0.60	0.28	0.58	0.41	0.52	0.56	0.58	0.34	0.37	0.52	0.53
	InDegreeBirth	0.24	-0.20	-0.08	0.19	0.27	0.24	0.27	0.23	0.29	0.08	0.25	0.22	0.86	1.00	0.76	0.86	0.86	0.82	-0.01	-0.13	-0.05	-0.07	-0.01	-0.16	0.53	0.54	0.46	0.50	0.55	0.44	0.52	0.46	0.48	0.49	0.51	0.41	0.30	0.48	0.49
	InDegreeEnd	0.19	-0.16	0.01	0.17	0.22	0.20	0.25	0.22	0.24	-0.01	0.26	0.16	0.77	0.76	1.00	0.77	0.78	0.82	0.01	-0.09	-0.04	-0.02	0.02	-0.13	0.47	0.36	0.60	0.45	0.48	0.41	0.45	0.35	0.53	0.45	0.45	0.39	0.30	0.39	0.40
In Degree	InDegreeAvg	0.27	-0.23	-0.06	0.20	0.29	0.26	0.33	0.28	0.35	0.07	0.32	0.25	0.96	0.86	0.77	1.00	0.98	0.74	0.02	-0.14	-0.07	-0.03	0.02	-0.17	0.57	0.35	0.43	0.55	0.59	0.31	0.55	0.44	0.49	0.54	0.54	0.39	0.33	0.51	0.52
	InDegreeMax	0.29	-0.25	-0.07	0.21	0.31	0.28	0.35	0.31	0.36	0.07	0.34	0.26	0.99	0.86	0.78	0.98	1.00	0.72	0.03	-0.14	-0.07	-0.02	0.03	-0.18	0.59	0.36	0.43	0.54	0.60	0.29	0.58	0.43	0.50	0.56	0.57	0.36	0.35	0.52	0.53
	InDegreeMin	0.15	-0.12	-0.03	0.13	0.18	0.16	0.18	0.14	0.20	0.02	0.16	0.13	0.70	0.82	0.82	0.74	0.72	1.00	-0.01	-0.14	-0.03	-0.05	0.00	-0.15	0.46	0.47	0.55	0.48	0.49	0.56	0.43	0.43	0.44	0.43	0.42	0.50	0.23	0.42	0.44
	OutDegreeDG	0.18	-0.13	-0.01	0.09	0.12	0.12	0.22	0.18	0.11	0.04	0.22	0.18	0.03	-0.01	0.01	0.02	0.03	-0.01	1.00	0.77	0.73	0.91	0.97	0.63	0.66	0.59	0.49	0.62	0.64	0.51	0.39	0.28	0.38	0.39	0.39	0.27	0.36	0.56	0.58
	OutDegreeBirth	0.02	0.03	0.03	0.01	-0.01	-0.01	0.02	0.00	-0.05	-0.02	0.02	0.02	-0.13	-0.13	-0.09	-0.14	-0.14	-0.14	0.77	1.00	0.74	0.81	0.79	0.84	0.45	0.69	0.47	0.49	0.46	0.58	0.12	0.27	0.16	0.14	0.13	0.26	0.20	0.36	0.36
	OutDegreeEnd	0.04	-0.04	0.12	0.02	0.01	0.02	0.06	0.05	0.03	-0.11	0.06	-0.02	-0.07	-0.05	-0.04	-0.07	-0.07	-0.03	0.73	0.74	1.00	0.75	0.75	0.90	0.43	0.57	0.66	0.48	0.44	0.70	0.22	0.16	0.39	0.22	0.22	0.28	0.21	0.30	0.30
Out Degree	OutDegreeAvg	0.14	-0.08	-0.02	0.06	0.07	0.07	0.16	0.13	0.07	0.04	0.15	0.14	-0.01	-0.07	-0.02	-0.03	-0.02	-0.05	0.91	0.81	0.75	1.00	0.94	0.72	0.56	0.57	0.48	0.62	0.58	0.53	0.32	0.23	0.34	0.33	0.33	0.26	0.29	0.51	0.52
	OutDegreeMax	0.16	-0.11	-0.02	0.10	0.11	0.11	0.20	0.17	0.10	0.04	0.20	0.16	0.03	-0.01	0.02	0.02	0.03	0.00	0.97	0.79	0.75	0.94	1.00	0.65	0.64	0.61	0.51	0.64	0.65	0.53	0.39	0.30	0.39	0.39	0.39	0.31	0.33	0.56	0.57
	OutDegreeMin	-0.01	0.03	0.12	-0.02	-0.06	-0.05	-0.04	-0.05	-0.07	-0.11	-0.03	-0.08	-0.17	-0.16	-0.13	-0.17	-0.18	-0.15	0.63	0.84	0.90	0.72	0.65	1.00	0.34	0.55	0.57	0.42	0.35	0.68	0.05	0.13	0.15	0.06	0.06	0.24	0.16	0.23	0.23
	DegreeDG	0.31	-0.23	-0.05	0.17	0.25	0.23	0.36	0.30	0.29	0.08	0.34	0.28	0.60	0.53	0.47	0.57	0.59	0.46	0.66	0.45	0.43	0.56	0.64	0.34	1.00	0.75	0.72	0.90	0.97	0.61	0.49	0.39	0.48	0.47	0.48	0.35	0.43	0.75	0.78
	DegreeBirth	0.14	-0.08	0.01	0.09	0.13	0.12	0.16	0.13	0.15	0.00	0.16	0.11	0.36	0.54	0.36	0.35	0.36	0.47	0.59	0.69	0.57	0.57	0.61	0.55	0.75	1.00	0.74	0.77	0.77	0.82	0.35	0.44	0.43	0.35	0.35	0.42	0.33	0.55	0.56
Total Degree	DegreeEnd	0.14	-0.12	0.12	0.09	0.11	0.11	0.17	0.15	0.15	-0.11	0.17	0.05	0.43	0.46	0.60	0.43	0.43	0.55	0.49	0.47	0.66	0.48	0.51	0.57	0.72	0.74	1.00	0.75	0.74	0.87	0.38	0.33	0.50	0.38	0.38	0.40	0.35	0.50	0.51
	DegreeAvg	0.26	-0.19	-0.03	0.14	0.21	0.19	0.29	0.24	0.26	0.04	0.28	0.22	0.54	0.50	0.45	0.55	0.54	0.48	0.62	0.49	0.48	0.62	0.64	0.42	0.90	0.77	0.75	1.00	0.93	0.71	0.43	0.38	0.43	0.43	0.42	0.38	0.39	0.72	0.74
	DegreeMax	0.29	-0.22	-0.05	0.17	0.25	0.23	0.34	0.28	0.29	0.07	0.32	0.25	0.60	0.55	0.48	0.59	0.60	0.49	0.64	0.46	0.44	0.58	0.65	0.35	0.97	0.77	0.74	0.93	1.00	0.64	0.50	0.41	0.49	0.48	0.49	0.38	0.41	0.75	0.77
	DegreeMin	0.08	-0.06	0.12	0.04	0.06	0.05	0.08	0.05	0.09	-0.11	0.08	0.00	0.28	0.44	0.41	0.31	0.29	0.56	0.51	0.58	0.70	0.53	0.53	0.68	0.61	0.82	0.87	0.71	0.64	1.00	0.29	0.34	0.39	0.29	0.28	0.42	0.31	0.43	0.44
	VBDG	0.26	-0.18	-0.12	0.27	0.33	0.31	0.34	0.29	0.28	0.13	0.33	0.29	0.58	0.52	0.45	0.55	0.58	0.43	0.39	0.12	0.22	0.32	0.39	0.05	0.49	0.35	0.38	0.43	0.50	0.29	1.00	0.73	0.95	0.97	0.99	0.64	0.40	0.38	0.38
	VBBirth	0.14	-0.05	-0.12	0.34	0.31	0.29	0.23	0.17	0.09	0.13	0.22	0.21	0.41	0.46	0.35	0.44	0.43	0.43	0.28	0.27	0.16	0.23	0.30	0.13	0.39	0.44	0.33	0.38	0.41	0.34	0.73	1.00	0.65	0.79	0.74	0.94	0.20	0.41	0.40
Vertex Betweeness	VBEnd	0.23	-0.12	-0.12	0.32	0.35	0.34	0.34	0.30	0.25	0.14	0.32	0.29	0.52	0.48	0.53	0.49	0.50	0.44	0.38	0.16	0.39	0.34	0.39	0.15	0.48	0.43	0.50	0.43	0.49	0.39	0.95	0.65	1.00	0.93	0.96	0.70	0.38	0.35	0.36
	VBAvg	0.23	-0.16	-0.12	0.29	0.31	0.29	0.32	0.28	0.24	0.13	0.30	0.28	0.56	0.49	0.45	0.54	0.56	0.43	0.39	0.14	0.22	0.33	0.39	0.06	0.47	0.35	0.38	0.43	0.48	0.29	0.97	0.79	0.93	1.00	0.98	0.72	0.37	0.37	0.38
	VBMax	0.25	-0.17	-0.12	0.28	0.32	0.31	0.33	0.29	0.27	0.13	0.32	0.29	0.58	0.51	0.45	0.54	0.57	0.42	0.39	0.13	0.22	0.33	0.39	0.06	0.48	0.35	0.38	0.42	0.49	0.28	0.99	0.74	0.96	0.98	1.00	0.66	0.40	0.38	0.38
	VBMin	0.10	0.01	-0.11	0.37	0.32	0.30	0.18	0.13	0.03	0.12	0.19	0.18	0.34	0.41	0.39	0.39	0.36	0.50	0.27	0.26	0.28	0.26	0.31	0.24	0.35	0.42	0.40	0.38	0.38	0.42	0.64	0.94	0.70	0.72	0.66	1.00	0.16	0.42	0.43
Clustering Coefficient	CCAvg	0.11	-0.14	0.06	0.15	0.20	0.19	0.19	0.17	0.14	-0.05	0.17	0.13	0.37	0.30	0.30	0.33	0.35	0.23	0.36	0.20	0.21	0.29	0.33	0.16	0.43	0.33	0.35	0.39	0.41	0.31	0.40	0.20	0.38	0.37	0.40	0.16	1.00	0.31	0.31
Edge Betweeness	EBCDG	0.29	-0.23	-0.04	0.01	0.08	0.06	0.28	0.23	0.21	0.06	0.26	0.19	0.52	0.48	0.39	0.51	0.52	0.42	0.56	0.36	0.30	0.51	0.56	0.23	0.75	0.55	0.50	0.72	0.75	0.43	0.38	0.41	0.35	0.37	0.38	0.42	0.31	1.00	0.95
	EBCAvg	0.31	-0.21	-0.06	0.03	0.10	0.08	0.31	0.25	0.22	0.09	0.29	0.23	0.53	0.49	0.40	0.52	0.53	0.44	0.58	0.36	0.30	0.52	0.57	0.23	0.78	0.56	0.51	0.74	0.77	0.44	0.38	0.40	0.36	0.38	0.38	0.43	0.31	0.95	1.00

Fig. 1 Kendall Correlation of all measures around the evolution of tables given the presence foreign keys

Finally, we classify tables with respect to different characteristics [2, 3]. Concerning table *survival*, we classify tables as *survivors* (present in the last version of the schema history that we study) or *dead* (otherwise). Concerning *update activity*, we classify tables in three groups: (a) *rigid*, with no updates at all in their lifetime, (b) quiet, if they have sustained less than 5 updates or an Average Transitional Update (ATU) rate less than 10%, and (c) *active,* if they have more than five updates and ATU higher than 10%. The Cartesian Product of *Survival* x *Activity* (resulting in 6 possible values) is referred to as *LifeAndDeath* (LAD) class.

		SurvivalClass	ActivityClass	LifeDeathClass
In Degree	InDegreeDG	0.08	**0.34**	**0.27**
	InDegreeBirth	0.08	0.25	0.22
	InDegreeEnd	-0.01	0.26	0.16
	InDegreeAvg	0.07	0.32	0.25
	InDegreeMax	0.07	**0.34**	0.26
	InDegreeMin	0.02	0.16	0.13
Out Degree	OutDegreeDG	0.04	0.22	0.18
	OutDegreeBirth	-0.02	0.02	0.02
	OutDegreeEnd	-0.11	0.06	-0.02
	OutDegreeAvg	0.04	0.15	0.14
	OutDegreeMax	0.04	0.20	0.16
	OutDegreeMin	-0.11	-0.03	-0.08
Total Degree	DegreeDG	0.08	**0.34**	**0.28**
	DegreeBirth	0.00	0.16	0.11
	DegreeEnd	-0.11	0.17	0.05
	DegreeAvg	0.04	0.28	0.22
	DegreeMax	0.07	0.32	0.25
	DegreeMin	-0.11	0.08	0.00
Vertex Betweeness	VBDG	**0.13**	**0.33**	**0.29**
	VBBirth	**0.13**	0.22	0.21
	VBEnd	**0.14**	0.32	**0.29**
	VBAvg	**0.13**	0.30	**0.28**
	VBMax	**0.13**	0.32	**0.29**
	VBMin	0.12	0.19	0.18
Clustering Coefficient	CCAvg	-0.05	0.17	0.13
Edge Betweeness	EBCDG	0.06	0.26	0.19
	EBCAvg	0.09	0.29	0.23

Fig. 2 Kendall correlation of Survival and Activity Classes with graph-theoretic measures of tables

Again, given this space of possibilities, it is still too hard to decide which measures are most promising to work with in order to discover interesting patterns. We have, thus, computed the Kendall correlation of all the measures we have in our possession to decide where the maximum correlations are located. The Kendall correlations that we have measured were consistent in all data sets, so, in Fig. 1 we depict the average value by averaging over all six of the datasets we have used, to give the general picture of the situation. We use green for all correlations that exceed 0.75 and red for all correlations that go under 0.75.

Practically speaking, there are very strong intra-correlations within the different variants of the same measure. So, it is sufficient to pick one measure from each family for our studies. Measures belonging to different families are moderately related to each other, with the respective values rarely exceeding .40. The only case of a strong correlation is between total degree and edge betweenness that exceeds 0.75. Overall, a measure's value at the Diachronic Graph typically has stronger correlations to the rest of the measures than any other of the candidate measures. Max value comes second and average value third. This is typical for all measures, although the differences are not very big, which practically means that if someone monitors the history of a schema, any of these measures is good enough.

As it is impossible to show all correlations of graph-theoretic measures to survival and activity, we resort to show a small fragment of the most important ones and in Fig. 2 we highlight the top 3 values of Kendall correlation for survival, activity and their combination. Survival is pretty much uncorrelated to graph metrics, with its Kendall value very close to zero (as we shall see later, there are differentiations from this general rule but for small sets of tables). Activity and LAD Classes, show a modest correlation to the graph metrics. In all occasions, it appears that the value of the Diachronic graph is the top value in all categories, followed by Max and Avg. Total Degree and Vertex Betweenness show the greater promise, with In-Degree following next. *The overall situation however, does not show clear signs of strong correlations, so, there is a need for an in-depth look in all categories.*

Given the profiles summarized in Fig. 2, as well as the detailed profile of each individual data set, we have decided that *the Total Degree of a table at the Diachronic Graph, DegreeDG, is both a simple measure*, that minimizes (as much as possible) the amount of tables having zero as value (compared to in-degree or out-degree, for example), and, at the same time, *the most promising measure to research on its relationship with survival and activity.* This explains why the paper has focused to *DegreeDG*.

1.2 Background statistics on the correlation of tables and foreign keys

As one can see in Fig. 3, there are two types of breakdown of nodes per degree. In 4 of the 6 data sets, we see a *monotone decrease* of the percentage of participating nodes, as the degree increases. This involves all the Computational Resource Toolkits (Castor, Egee) and the CMSs (Slashcode and Zabbix). The tables that have no relationship to foreign keys constitute the majority in all of them (absolute in 3 out of 4, and with large percentages). The two scientific datasets, on the other hand, enforce the inter-relationship of the tables via foreign keys. Biosql is practically having a constant distribution of tables to degrees, whereas Atlas has a *battleship pattern*, with a monotone increase in the first 3 values and a sharp decrease (occasionally with a long right tail). In all the bar charts, we have grouped the degrees of value higher or equal to 4 in a single group, as most values in this group do not exceed 2 tables per degree.

Degree DG	0	1	2	3	>=4	Total
atlas	11	25	35	7	10	88
biosql	2	11	13	9	10	45
castor	75	8	6	2	0	91
egee	6	2	2	2	0	12
slashcode	90	21	7	1	7	126
zabbix	23	15	13	2	5	58

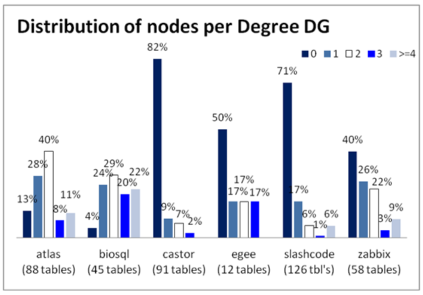


Fig. 3 Breakdown of nodes per Degree at the Diachronic Graph

Atlas 88 tables — outDegreeDG (columns) × inDegreeDG (rows)

inDegreeDG	0	1	2	[4-10]	
0	13%	13%	28%	1%	55%
1	16%		2%	2%	20%
2	11%	1%			13%
[3-7]	9%	2%	1%		13%
	49%	16%	32%	3%	100%

SlashCode 126 tables — outDegreeDG (columns) × inDegreeDG (rows)

inDegreeDG	0	1	2	[3-5]	
0	71%	14%	5%		90%
1	2%	1%			3%
2		1%	1%	1%	2%
[4-14]	2%		1%	2%	4%
	75%	16%	6%	2%	100%

Biosql 45 tables — outDegreeDG (columns) × inDegreeDG (rows)

inDegreeDG	0	1	2	[3-4]	
0	4%	18%	24%	20%	67%
1	7%	2%	2%	2%	13%
2		2%		2%	4%
[3-17]	4%	4%	4%	2%	16%
	16%	27%	31%	27%	100%

Zabbix 58 tables — outDegreeDG (columns) × inDegreeDG (rows)

inDegreeDG	0	1	2	3	
0	40%	19%	12%	2%	72%
1	7%	5%			12%
2	5%		2%		7%
[3-8]	3%	2%	3%		9%
	55%	26%	17%	2%	100%

Fig. 4 How do InDegreeDG and OutDegreeDG co-occur. The shaded area of the cells corresponds to nodes with total degree higher than 2.

In Figure 4, we depict how the In- and Out-Degree at the Diachronic Graph co-occur in 4 of the 6 data sets. The 2 data sets missing resemble the ones at the right-hand side of the Figure, albeit with maximum out-degree of 2 and just one table in Castor with in-degree higher than 2. We are primarily interested to exploit this relationship in *understanding high-degree nodes of the graph* (i.e., nodes with value higher than 2 for their *total* degree). We have 3 fundamental observations: (a) the *number* of tables having out-degree higher than zero is slightly bigger than the number of tables having in-degree higher than zero, (b) the probability of a table having both a non-zero in-degree and a non-zero out-degree is quite small, and, (c) with the exception of Atlas, the *range* of degrees is much higher for in-degrees (i.e., there do exist -few- tables with many foreign keys pointing to them) rather than out-degrees (i.e., tables with a large number of outgoing foreign keys).The last observation signifies that, in the subsequent deliberations on tables with high total degree, this population is mainly distinguished by its high in-degree rather than its high out-degree at the Diachronic Graph.

1.3 Details on Total Degree and its relationship to Survival

Both the Last Known Version (LKV) and the Duration of the tables (Fig. 5) demonstrate the same behavior over DegreeDG, summarized as a Gamma pattern: small degrees exhibit all kind of durations and last known versions, whereas high degrees miss the highest value (equivalently: do not survive) very seldom. Fig. 5 depicts the Gamma pattern. Remember that the scatterplots are painted with transparency: intense colors show high concentration of points at the same x,y coordinates, whereas lighter color means a single (or few) point(s). Durations follow the same pattern too.

Although not depicted here, we have found that the same pattern repeats itself for in-degrees too, but not exactly for out-degrees. The latter is mainly due to the absence of out-degrees of high value. Remember that tables without degree are more than tables with in degree, with the later being able, though, to reach high values of in-degree.

1.4 Details on Total Degree and its relationship to Activity

We demonstrate here the statistical details on how Total Degree correlates to Activity.We grouped tables in a 3x3 contingency table with the groups formed as follows. Concerning activity, we keep the 3 values (rigid, quiet, active). Concerning total degree at DG, we group tables in (a) 0 degree, (b) degree within 1 and 2, and (c) degree higher than 2. Figure 6 presents the contingency tables for all the data sets.

1.5 The interesting case of high-degree active tables

Research Question: what are the characteristics of the tables that are both active and have a high total degree at the Diachronic Graph? The paper sets apart exactly these tables that come with both a high total degree at the Diachronic Graph and high activity. What is it that characterizes this activity?

In Figure 7, we depict a numerical assessment of the characteristics of these tables. An observing eye will notice the following properties:

- The tables whose high degree value is due to their high in-degree at the Diachronic Graph are typically more than the ones with high out-degree at the Diachronic Graph. The only exception is Biosql, due to its unique schema structure than involves long chains of foreign keys between tables.
- Not shown in the figure is that only 2 of the overall 33 tables in the 6 data sets belong both to the high out-degree and the high in-degree category; at the same time, 7 tables belong to neither of the above and only to the bottom part of the figure (meaning that it is the combination of in and out degree that makes them exceed the total degree value of 2 at the Diachronic Graph).
- The statistics in all cases demonstrate that *a major reason for the high update nature of these tables is the perfective maintenance performed to these tables*. This is demonstrated by two facts: (a) the really high percentage of deleted attributes, for *all* data sets and (b) the fairly high rate of data type updates, again for *all* data sets. Taking into account that most tables live a rigid or quiet life, both

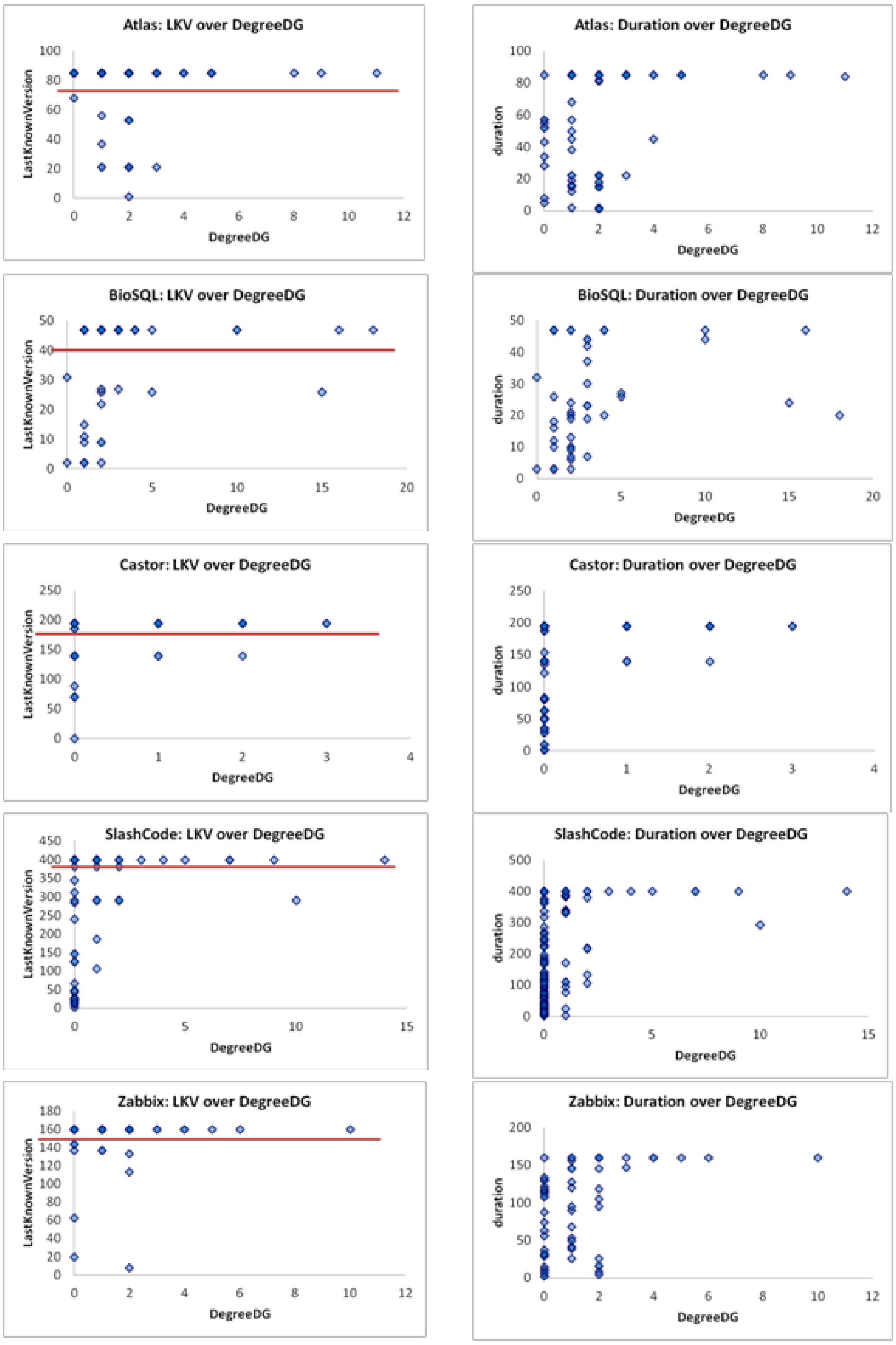


Fig. 5 Relationship of DegreeDG with Last Known Version and Duration.

Atlas (88 tables)

TotDegDG	Rigid	Quiet	Active	Σ
0	3%	7%	2%	*13%*
1-2	17%	34%	17%	*68%*
>2	0%	8%	11%	*19%*
Σ	*20%*	*49%*	*31%*	*100%*

BioSQL (45 tables)

TotDegDG	Rigid	Quiet	Active	Σ
0	4%	0%	0%	*4%*
1-2	24%	22%	7%	*53%*
>2	7%	7%	29%	*42%*
Σ	*36%*	*29%*	*36%*	*100%*

Castor (91 tables)

TotDegDG	Rigid	Quiet	Active	Σ
0	55%	26%	1%	*82%*
1-2	8%	7%	1%	*15%*
>2	0%	1%	1%	*2%*
Σ	*63%*	*34%*	*3%*	*100%*

Slashcode (126 tables)

TotDegDG	Rigid	Quiet	Active	Σ
0	40%	28%	4%	*71%*
1-2	3%	14%	5%	*22%*
>2	0%	2%	4%	*6%*
Σ	*43%*	*44%*	*13%*	*100%*

Zabbix (58 tables)

TotDegDG	Rigid	Quiet	Active	Σ
0	22%	16%	2%	*40%*
1-2	21%	28%	0%	*48%*
>2	0%	9%	3%	*12%*
Σ	*43%*	*52%*	*5%*	*100%*

Fig. 6 Breakdown of tables with respect to their activity profile and their Total Degree at the Diachronic Graph

Pct over Total Amt of Updates...			... for attributes not involved in FK's					...for attributes involved in FK's					... for all attributes in total				
Active, inDegDG > 2	#tables	Total Amt of Upd's	Injections	Ejections	Type Upd	Key Upd	Total	Injections	Ejections	Type Upd	Key Upd	Total	Injections	Ejections	Type Upd	Key Upd	Total
Atlas	5	59	25%	19%	53%	0%	97%	0%	3%	0%	0%	3%	25%	22%	53%	0%	100%
Biosql	4	57	37%	26%	11%	0%	74%	16%	11%	0%	0%	26%	53%	37%	11%	0%	100%
Castor	1	21	57%	38%	5%	0%	100%	0%	0%	0%	0%	0%	57%	38%	5%	0%	100%
Egee	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Slashcode	3	192	30%	21%	21%	4%	76%	6%	5%	9%	4%	24%	36%	26%	30%	8%	100%
Zabbix	1	31	71%	6%	10%	0%	87%	6%	6%	0%	0%	13%	77%	13%	10%	0%	100%
Active, outDegDG > 2	#tables	Total	Injections	Ejections	Type Upd	Key Upd	Total	Injections	Ejections	Type Upd	Key Upd	Total	Injections	Ejections	Type Upd	Key Upd	Total
Atlas	2	21	19%	0%	33%	0%	52%	19%	29%	0%	0%	48%	38%	29%	33%	0%	100%
Biosql	10	120	20%	17%	6%	2%	44%	22%	19%	0%	15%	56%	42%	36%	6%	17%	100%
Castor	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Egee	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Slashcode	2	153	27%	22%	18%	0%	67%	10%	6%	14%	3%	33%	38%	28%	31%	3%	100%
Zabbix	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Active, DegDG > 2	#tables	Total	Injections	Ejections	Type Upd	Key Upd	Total	Injections	Ejections	Type Upd	Key Upd	Total	Injections	Ejections	Type Upd	Key Upd	Total
Atlas	10	118	25%	15%	49%	0%	90%	3%	7%	0%	0%	10%	29%	22%	49%	0%	100%
Biosql	13	163	26%	21%	7%	1%	55%	19%	15%	0%	11%	45%	45%	36%	7%	12%	100%
Castor	1	21	57%	38%	5%	0%	100%	0%	0%	0%	0%	0%	57%	38%	5%	0%	100%
Egee	2	37	49%	22%	22%	3%	95%	5%	0%	0%	0%	5%	54%	22%	22%	3%	100%
Slashcode	5	287	38%	23%	6%	6%	73%	0%	0%	27%	0%	27%	38%	23%	33%	6%	100%
Zabbix	2	58	62%	21%	9%	0%	91%	3%	5%	0%	0%	9%	66%	26%	9%	0%	100%

Fig. 7 Breakdown of active tables with high degrees with respect to (i) high in-degree(top), high out-degree (middle), high total degree (bottom), (ii) changes for attributes not involved in foreign keys (second pillar), changes for attributes involved in foreign keys (third pillar), overall (rightmost pillar), and, (iii) attribute injections, ejections, data type change and primary key change (within each pillar)

these aspects are quite untypical. The deletions are mostly due to renames or do-undo actions (i.e., injection of an attribute at a certain version and ejection shortly after).

– Even if we subtract the percentage of ejections from the percentage of injection of attributes (i.e., even if we assume that all deletions undo an insertion), the remaining percentage of injections demonstrates that there is also *an increase in the information capacity of these tables*.
– For the two first, scientific, data sets, an extra observation is the high percentage of restructuring and updates in attributes concerning foreign keys. This is due to the particularity of the overall evolution of their schema: these two data sets have undergone too active schema restructuring (especially Biosql), with reference tables being deleted, or renamed, and their key attributes' data types being updated. This results in the need to update attributes of the referencing tables which are foreign keys.

Overall, we can argue that *the reason for the high update rate of the high-degree, active tables is an unusual combination of information capacity expansion and perfective maintenance, the latter being expressed via data type updates as well as do-undo actions*.

We take advantage of this discussion to demonstrate the Diachronic Graphs of our data sets, annotated with the location of high-degree active tables, in Figures 8, 9, and 10.

1.6 Details on the relationship of the Total Degree of a table with its Birth

In the paper we mention that, it is clear that (with the exception of BioSQL that has exactly the opposite behavior) *there is strong evidence that high degree tables are not born after the originating version of the database.*

Can we attribute the phenomenon to the nature of DegreeDG? *Is it possible that it is due to the fact that the tables are born with small degree, and progressively*

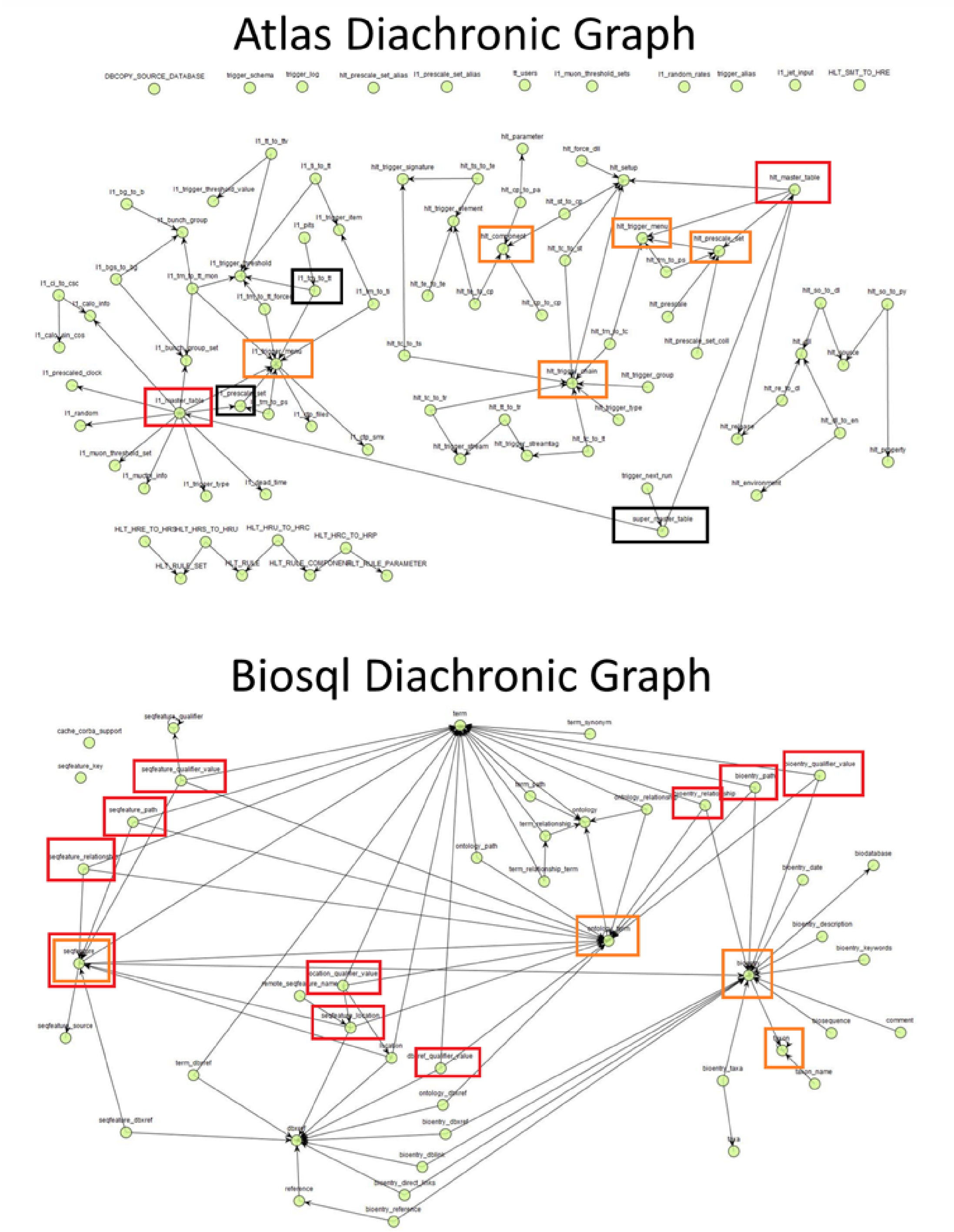


Fig. 8 Diachronic Graph of scientific datasets, annotated with "boxing" high-degree active tables: orange signifies high inDegreeDG, red high outDegreeDG and black neither

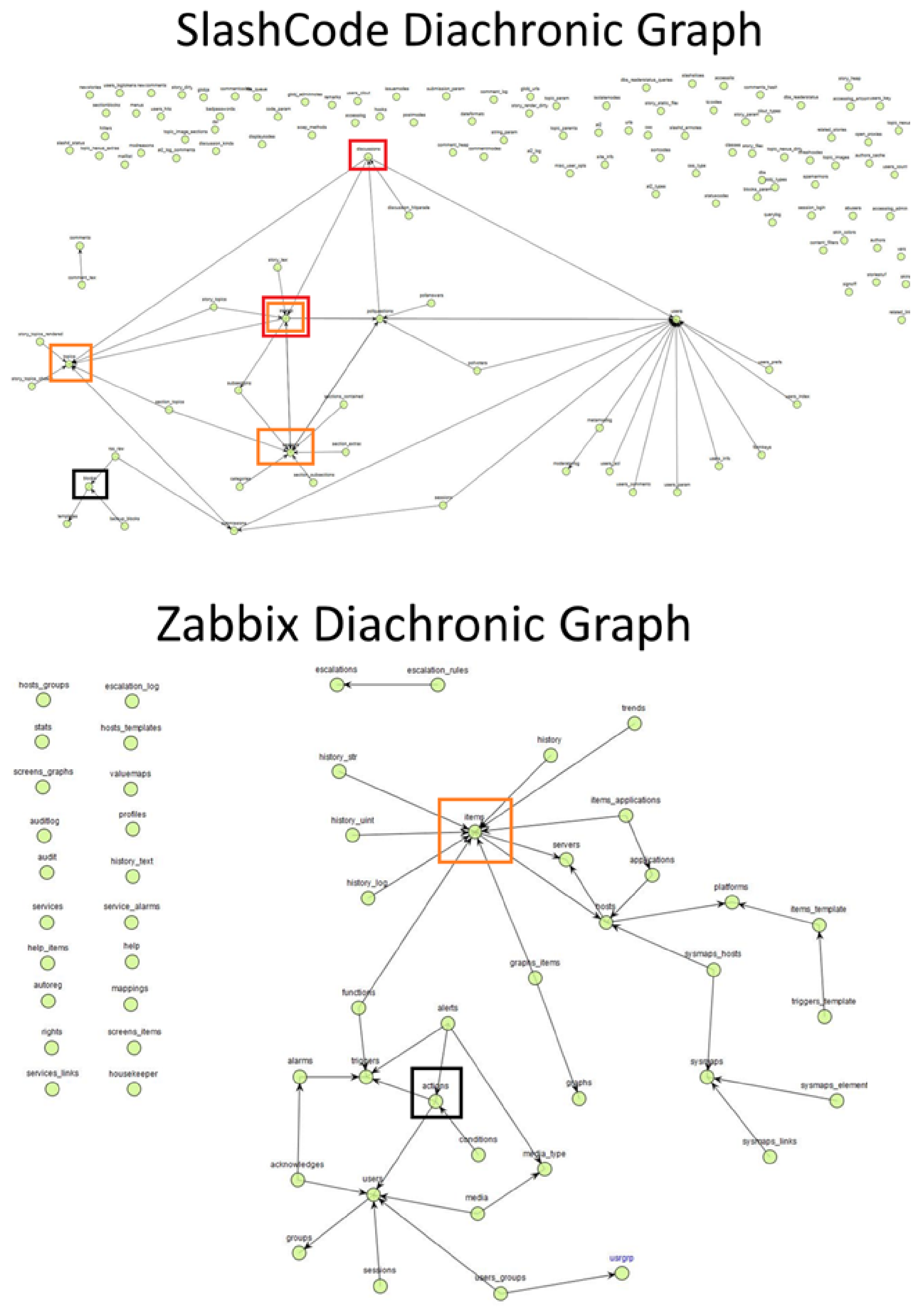


Fig. 9 Diachronic Graph of CMS datasets, annotated with "boxing" high-degree active tables: orange signifies high inDegreeDG, red high outDegreeDG and black neither

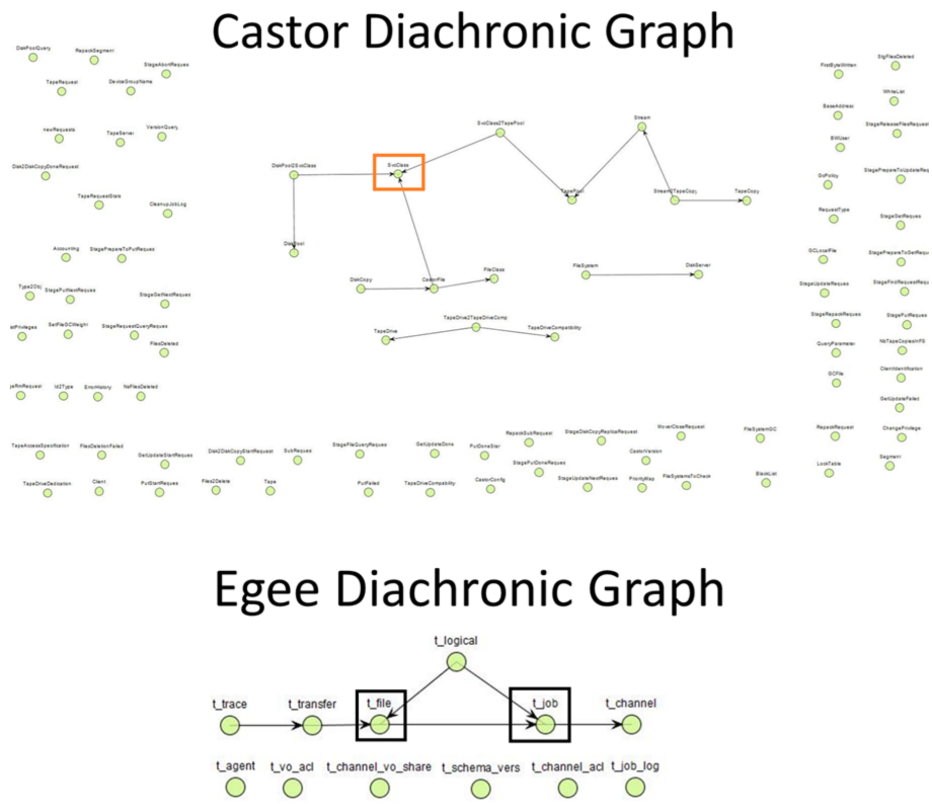


Fig. 10 Diachronic Graph of resource management datasets, annotated with "boxing" high-degree active tables: orange signifies high inDegreeDG, red high outDegreeDG and black neither

they built it up, over their lifetime? Although certain tables do scale-up in their degree, and most certainly, DegreeDG is quite larger than average or degree at birth (mainly due to the deletion of certain tables that remain in the DG), the answer is negative. In Fig. 11, we also depict the aforementioned alternatives: although both reach smaller values than DegreeDG, the shape of the scatterplots is consistent: any table with (any form of) total degree above 2, is pretty much born at version 0.

Also, we cannot attribute the phenomenon to the breakdown of tables to degrees (see for example Fig. 3 on how unrelated the percentage of tables belonging to a particular degree is to the range of birth values in Fig. 11).

1.7 Other metrics and their relationships to DegreeDG

We have also performed preliminary tests –in the form of scatterplots- on the relationships of other metrics with the total degree of a node at the Diachronic Graph. These involved metrics like schema scale-up, death, sum of updates and

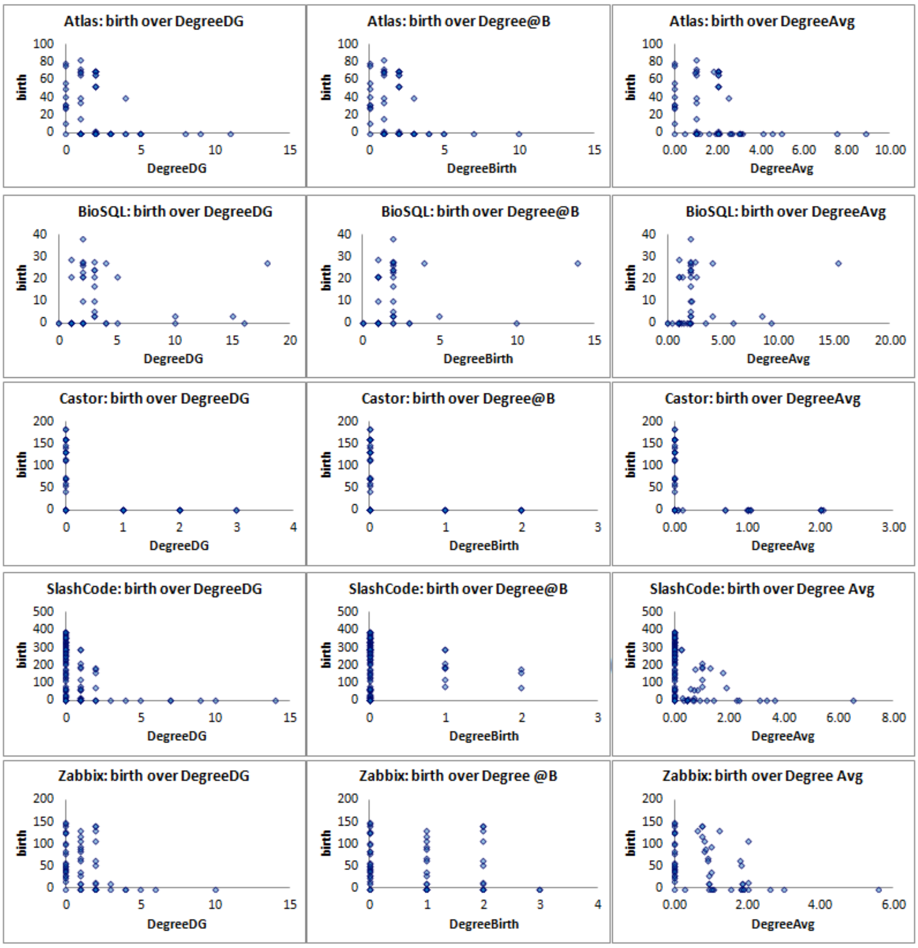


Fig. 11 Left: birth over degreeDG; center: over degree @ Birth; right: over average degree

ATU. The overall behavior does not seem to show anything more than following the general distribution of tables. Of course, further study might reveal hidden patterns that the preliminary inspection might have missed.

We have also researched in- and out-degrees in depth, as well as other measures like vertex betweenness, clustering coefficient, and edge betweenness, albeit in less depth. The discussion of these measures would double the size of this paper without much gain: our findings so far demonstrate that Total Degree at the Diachronic Graph is the best measure to obtain insights concerning the survival and the update activity of a table. Here too, it possible that further inspection reveals more findings that we might have missed.

References

1. Skoulis I, Vassiliadis P, Zarras AV (2015) Growing up with stability: How open-source relational databases evolve. Information Systems 53:363–385
2. Vassiliadis P, Zarras AV, Skoulis I (2015) How is life for a table in an evolving relational schema? birth, death and everything in between. In: Proceedings of 34th International Conference on Conceptual Modeling (ER 2015), Stockholm, Sweden, October 19-22, 2015, pp 453–466
3. Vassiliadis P, Zarras AV, Skoulis I (2017) Gravitating to rigidity: Patterns of schema evolution - and its absence - in the lives of tables. Information Systems 63:24–46