Воспроизведение результатов статьи в pygraphs.

Владимир Ивашкин

7 октября 2018 г.

2 Logarithmic vs. plain measures

Не ясно, в оригинале был RI или ARI. Если был ARI, то он на тот момент был неправильным. Привожу тут оба варианта

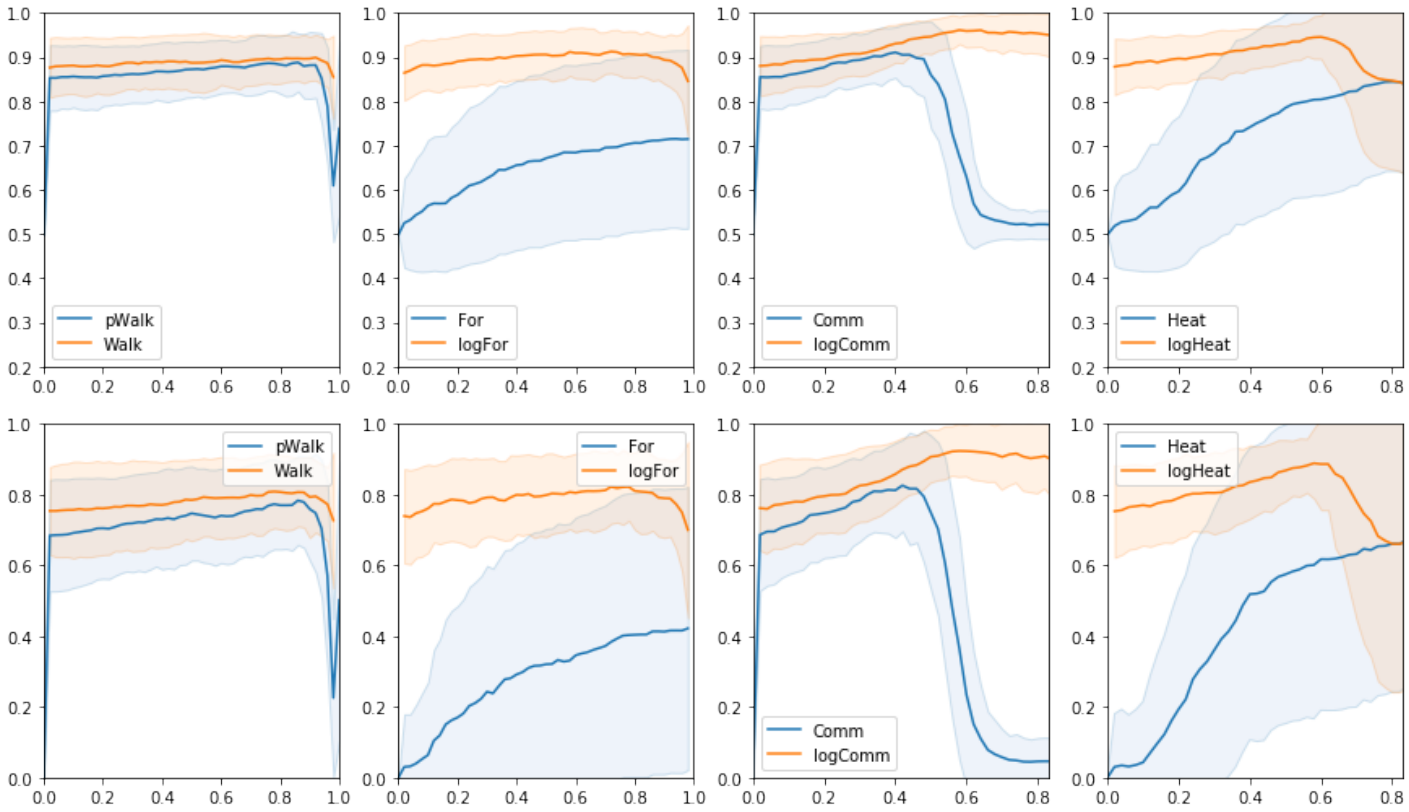


Рис. 1: $G(100, (2)0.2, 0.05)$, RI and ARI respectively

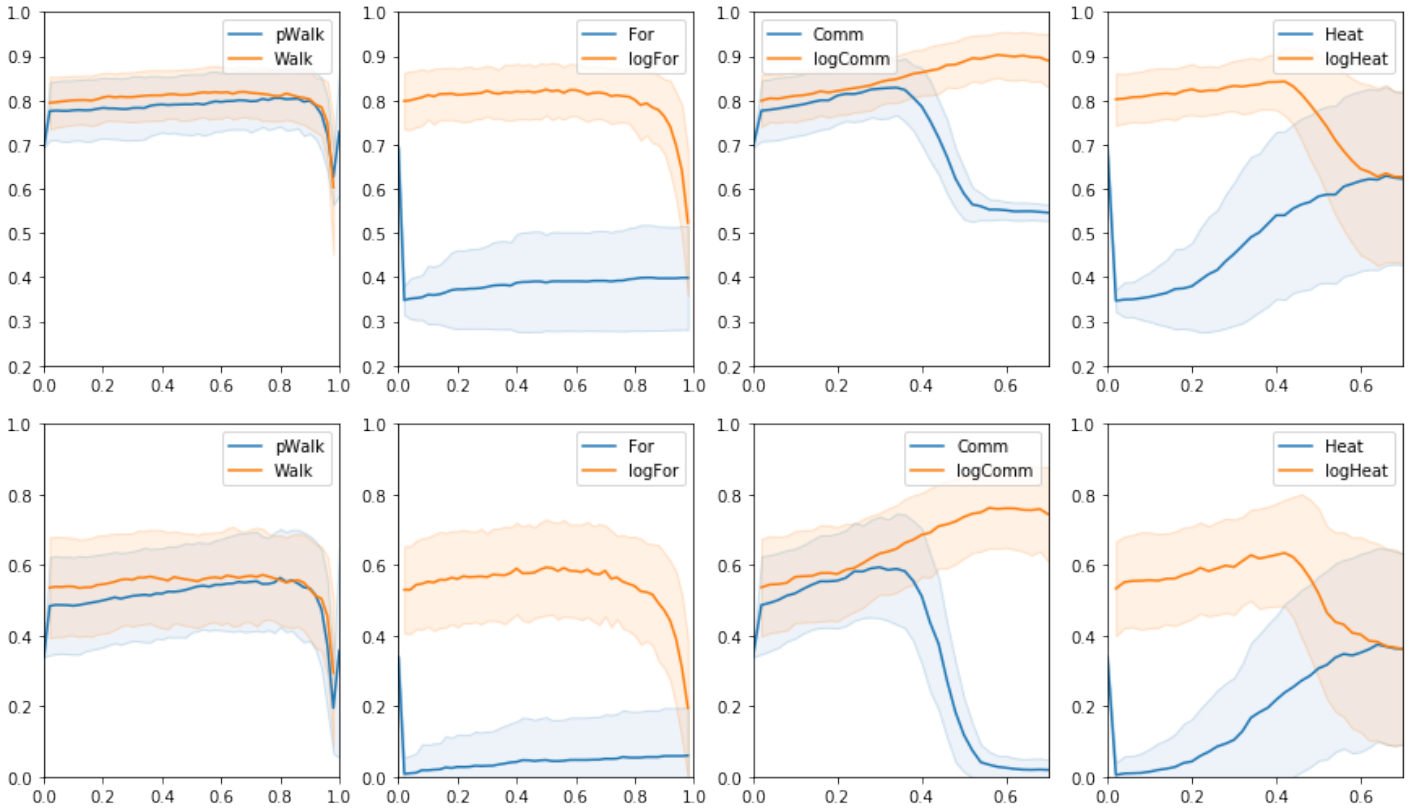


Рис. 2: $G(100, (3)0.3, 0.1)$, RI and ARI respectively

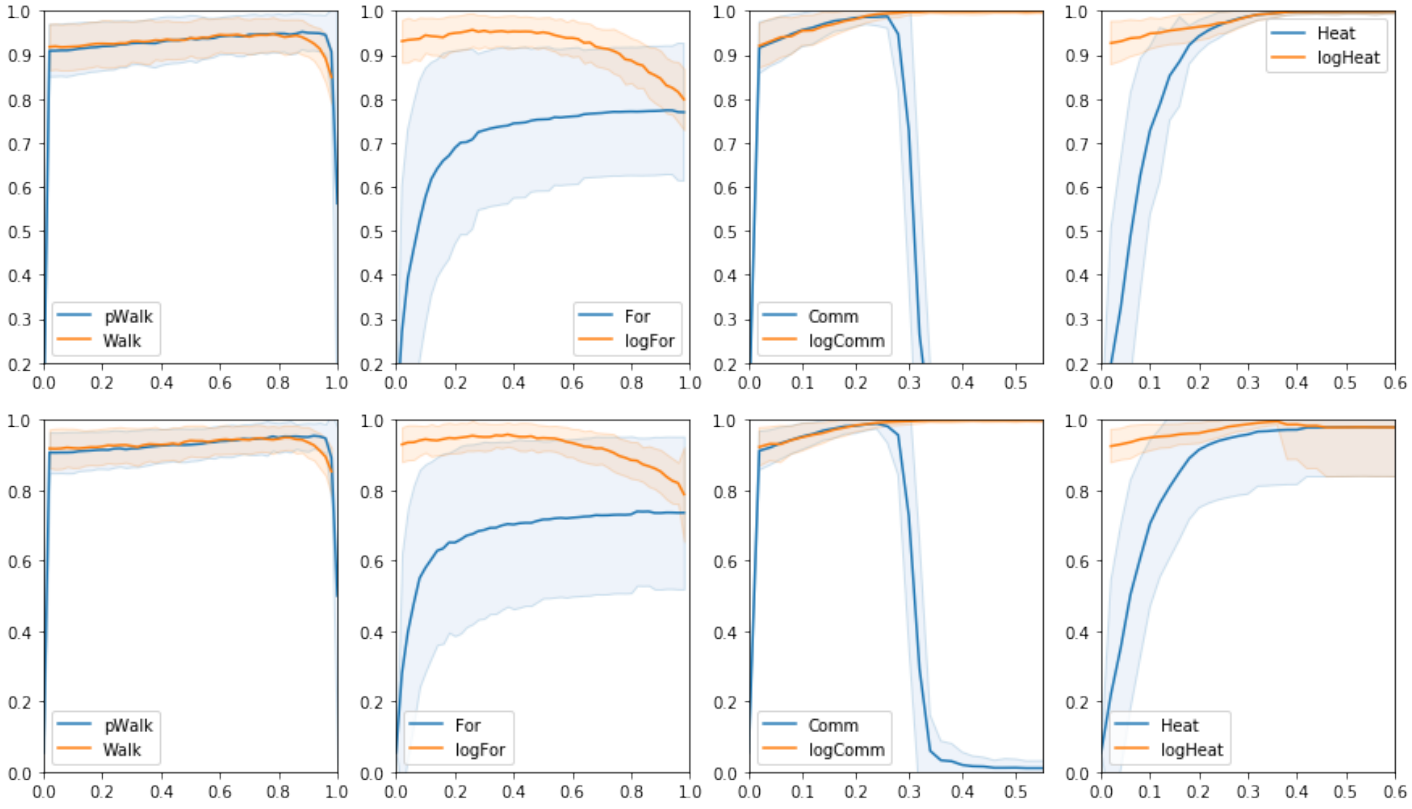


Рис. 3: $G(200, (2)0.3, 0.1)$, RI and ARI respectively

3 Competition by Copeland's score

Nodes	100	100	100	100	200	200	200	200	**Sum**
Classes	2	2	4	4	2	2	4	4	**of**
p_{out}	0.1	0.15	0.1	0.15	0.1	0.15	0.1	0.15	**scores**
logComm	10	512	406	-122	580	333	152	600	**2471**
Comm	4	185	86	448	244	297	442	246	**1952**
SCCT	10	287	188	148	289	238	76	458	**1694**
Heat	10	-310	86	448	136	332	442	-260	**884**
pWalk	-3	-41	86	448	-41	-106	442	-138	**647**
logHeat	4	67	-16	-294	202	332	-292	166	**169**
SCT	-6	51	-106	148	-39	69	76	-42	**151**
logFor	-8	33	-70	-298	3	-83	-262	50	**−635**
FE	0	-12	-104	-294	-97	-102	-294	-4	**−907**
For	-10	-560	86	448	-568	-546	442	-260	**−968**
RSP	-3	92	-132	-358	-107	-1	-336	-124	**−969**
Walk	4	20	-40	-316	-144	-221	-346	-98	**−1141**
SP-CT	-12	-324	-470	-406	-458	-542	-542	-594	**−3348**

Таблица 1: Optimal parameters

Nodes	100	100	100	100	200	200	200	200	**Sum**
Classes	2	2	4	4	2	2	4	4	**of**
p_{out}	0.1	0.15	0.1	0.15	0.1	0.15	0.1	0.15	**scores**
logComm	440	501	466	340	398	565	574	582	**3866**
SCCT	263	295	360	184	295	397	438	370	**2602**
Comm	109	149	106	120	198	60	168	158	**1068**
logHeat	236	59	80	32	391	11	148	98	**1055**
logFor	-23	57	148	116	-126	44	134	94	**444**
FE	-74	80	50	120	-30	30	38	52	**266**
Walk	-79	119	114	102	-84	-4	20	76	**264**
SCT	-27	27	4	-32	52	-6	36	30	**84**
pWalk	45	1	20	10	-62	-31	-10	26	**−1**
Heat	296	-322	-492	-445	386	249	-215	-472	**−1015**
RSP	-313	-117	-16	14	-338	-268	-280	-84	**−1402**
SP-CT	-482	-287	-250	0	-585	-460	-452	-352	**−2868**
For	-391	-562	-590	-561	-495	-587	-599	-578	**−4363**

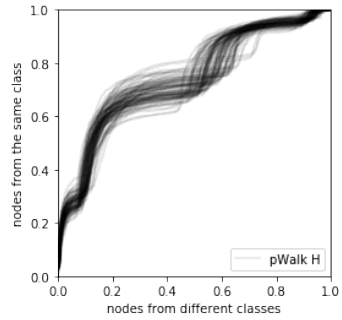
Таблица 2: 90th percentiles

4 Reject curves

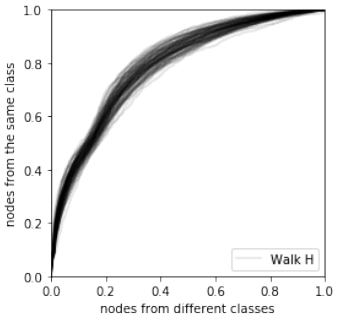
Measure (kernel)	$G(100, (2)0.3, 0.05)$ Opt. parameter, ARI	$G(100, (2)0.3, 0.1)$ Opt. parameter, ARI	$G(100, (2)0.3, 0.15)$ Opt. parameter, ARI
pWalk	0.93, 1.00	0.87, 0.91	0.73, 0.66
Walk	0.93, 1.00	0.67, 0.91	0.70, 0.65
For	0.60, 0.99	0.97, 0.51	0.40, 0.01
logFor	0.70, 1.00	0.40, 0.93	0.10, 0.68
Comm	0.33, 1.00	0.33, 0.98	0.30, 0.77
logComm	0.33, 1.00	0.47, **1.00**	0.57, **0.91**
Heat	0.37, 1.00	0.60, 0.87	0.73, 0.15
logHeat	0.37, 1.00	0.53, 0.99	0.37, 0.80
SCT	0.40, 1.00	0.57, 0.94	0.43, 0.72
SCCT	0.03, 1.00	0.57, 0.98	0.63, 0.80
RSP	0.97, 1.00	0.97, 0.93	0.97, 0.67
FE	0.90, 1.00	0.90, 0.91	0.87, 0.68
SP-CT	0.00, 0.99	0.03, 0.78	0.07, 0.49

Таблица 3: Optimal family parameters and the corresponding ARI's

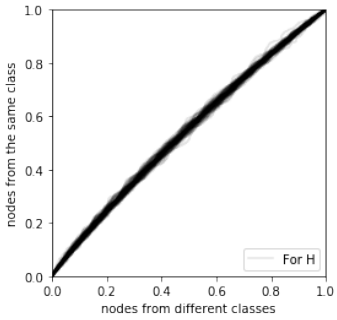
Ошибка была в том, что подобранные параметры из таблицы выше принадлежат к диапазону $[0, 1]$, а значит их нужно преобразовывать к диапазону, специфичному для конкретной метрики. Я же этого не делал. Вторая ошибка состояла в том, что я использовал тут близости вместо расстояний. Еще тогда, когда я строил их в прошлый раз, я заметил, что по близостям logComm совсем не обгоняет остальные меры, но по расстояниям эффект выраженный. Тут его тоже видно:



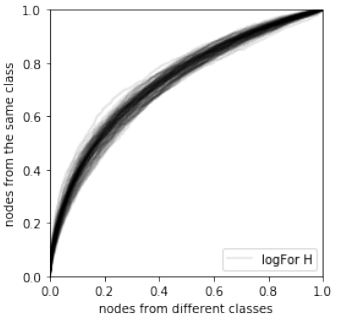
(a) pWalk



(b) Walk



(c) For



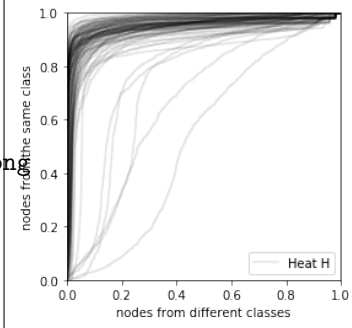
(d) logFor

pictures/4_Comm_alt.png

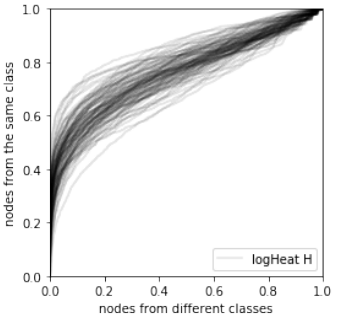
(e) Comm

pictures/4_logComm_alt.png

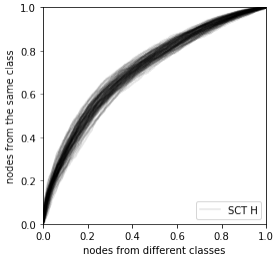
(f) logComm



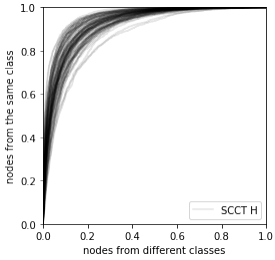
(g) Heat



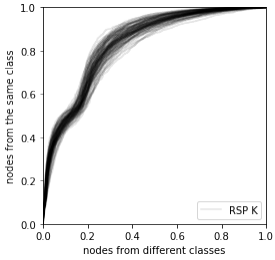
(h) logHeat



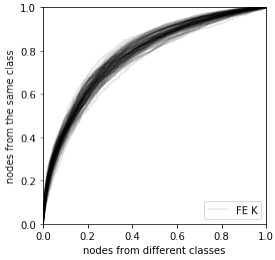
(i) SCT



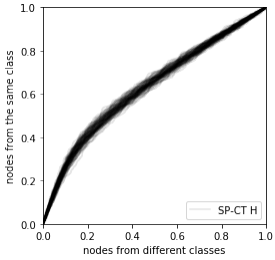
(j) SCCT



(k) RSP



(l) FE



(m) SP-CT

Рис. 4: Reject curves for the graph measures under study

pictures/4_all_alt.png

(a) All families

pictures/4_4best_alt.png

(a) All families

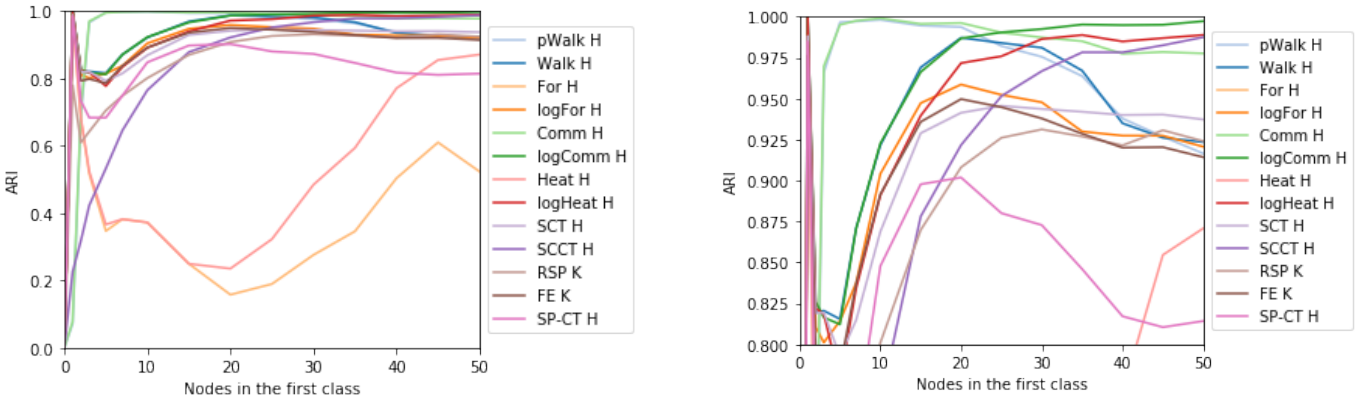
Рис. 5: Average reject curves

Здесь была проблема со взятием корня из Comm/logComm. А проблема была такая: если некоторые значения матрицы D при взятии корня превращаются в nan, то стандартная сортировка оставляет их на тех же позициях и

отдельно сортирует массив слева и справа от них. Получается кусочно-возрастающая функция, из которой потом получаются несколько маленьких reject curve вместо одной большой. Решение – фильтровать эти nan и сортировать без них. Раз такой эффект вообще возник, значит в матрице D иногда появляются отрицательные значения.

Можно подозревать внешний вид графика pWalk. Может быть, это связано с тем, как мы фиксируем параметр. Параметром считаем отскалированое в [0, 1] число, для каждого графа преобразуем его в зависимости от спектрального радиуса матрицы A $(param = t/\rho(A)), t \in [0, 1])$.

5 Graphs with classes of different sizes



(a) All families

(b) Leading families

Рис. 6: Graphs with two classes of different sizes: clustering with optimal parameter values

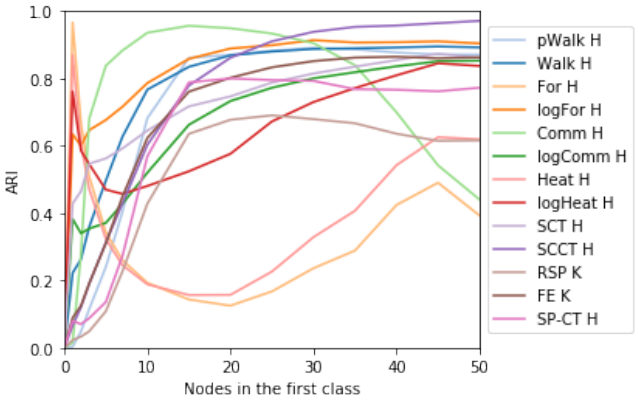
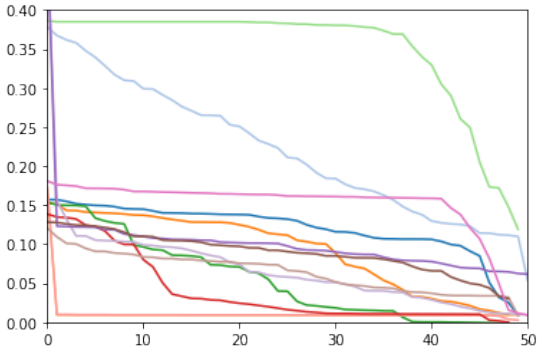


Рис. 7: Graphs with two classes of different sizes: random parameter values

$$P = \begin{pmatrix} 0.30 & 0.20 & 0.10 & 0.15 & 0.07 & 0.25 \\ 0.20 & 0.24 & 0.08 & 0.13 & 0.05 & 0.17 \\ 0.10 & 0.08 & 0.16 & 0.09 & 0.04 & 0.12 \\ 0.15 & 0.13 & 0.09 & 0.20 & 0.02 & 0.14 \\ 0.07 & 0.05 & 0.04 & 0.02 & 0.12 & 0.04 \\ 0.25 & 0.17 & 0.12 & 0.14 & 0.04 & 0.40 \end{pmatrix}.$$

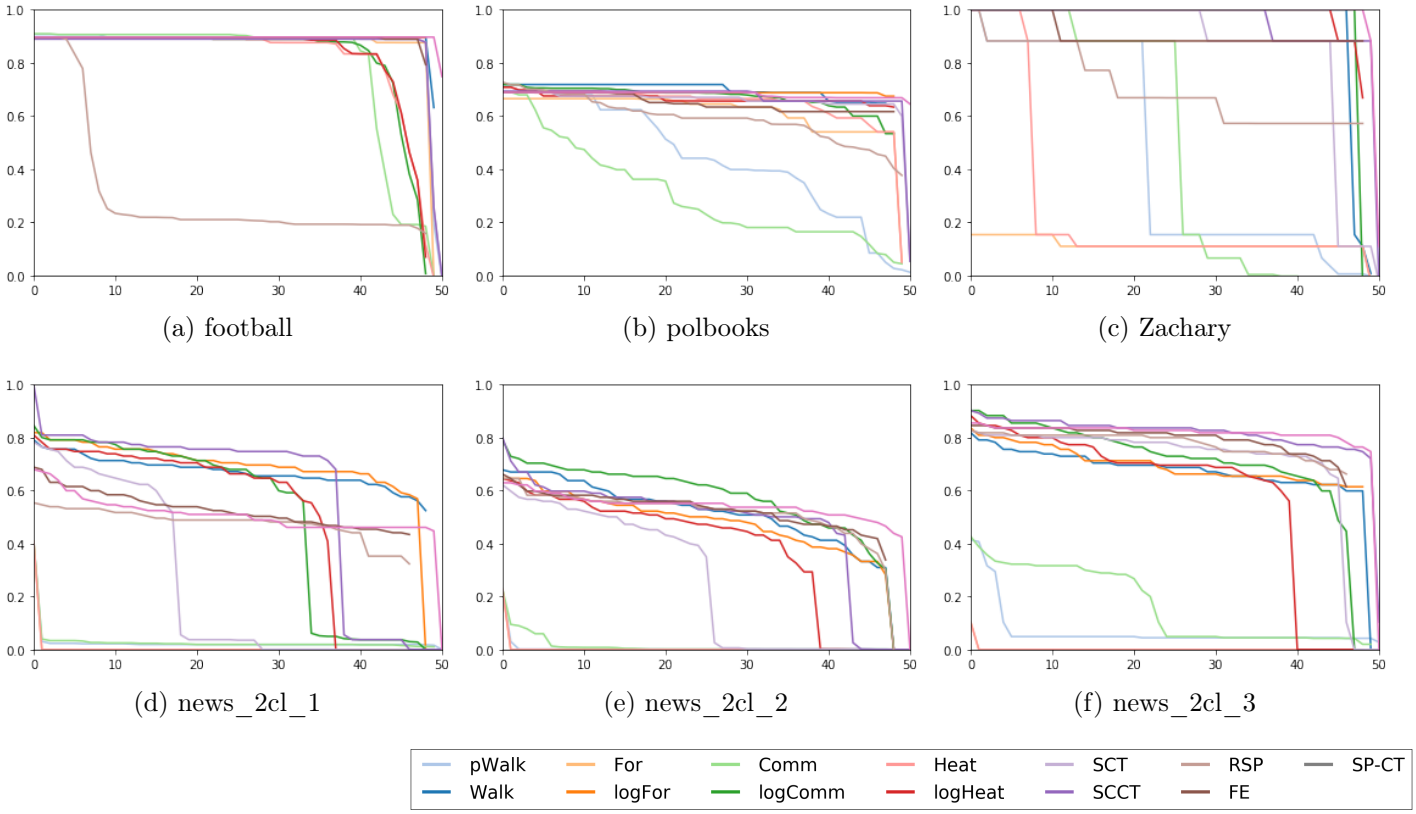


.45.45ARI

of various measure families on a structure with 6 classes

6 Cluster analysis on several classical datasets

Здесь ошибка была в том, что я зафиксировал число классов – 2, хотя в датасете football их 12. Теперь все похоже на статью:



(a) football (b) polbooks (c) Zachary

(d) news_2cl_1 (e) news_2cl_2 (f) news_2cl_3

Рис. 8: ARI of various measure families on classical datasets