User Study Summary for PieceStack

To further demonstrate the effectiveness of PieceStack, we conducted a comparative study with a benchmark system formed by a classical stacked graph. In the benchmark system, users can drag layers freely, place layers with respect to their interests, or arrange layers around their preferred baselines directly.

Twenty experienced computer users were recruited, including 13 males and 7 females. The average age of the participants was 23.5 (from 21 to 29, median 23). Six of them reported to had seen a stacked graph before, and one was familiar with time series visualization techniques. A compact dataset containing twelve layers (60 time points for each layer) with clear shape patterns was chosen for this user study. At the beginning of each user session, we provided a brief tutorial of both systems, and encouraged them to freely try both systems and to ask questions. The participants were then asked to finish five tasks with both systems, first the benchmark system and then PieceStack. Considering the complexity involved and time required to perform the tasks, the following simple yet representative tasks were selected:

T.1 Find the most contributive layer(s) for an aggregated bump;
T.2 Find the types of contributions (i.e., significantly positive, significantly negative, or trivial) of all layers to an overall bump;
T.3 For two aggregated bumps, find the layers that have the same type of contributions;
T.4 Find the layer(s) that contributes the most to the overall aggregated shape on the whole time period;
T.5 For a given layer, find layers that are similar to it at least in some time intervals.

These tasks were designed for verifying our system's capability of solving the tasks problems we summarized in the Design Consideration: T.1, T.2 and T.4 all belong to *layer versus aggregation* problems, with T.1 and T.4 testifying the similarity between a layer and the aggregation, and T.2 the generation of aggregation patterns. Specifically, T.1 emphasizes local features by focusing on one time point, and T.4 generalizes it to understand how users observe the global features. T.3 essentially compares the causes of two aggregated features and could be categorized as aggregation versus aggregation. T.5 is for the comparison between layers. Though we recognize it as the cornerstone for all other tasks involving aggregations, it is placed at the end since the uncertainness of similar interval's position complicates the comparison procedure.

We analyzed the answers and the time that the participants spent on each task with both systems. Overall, our system guaranteed improvements in accuracy and efficiency. While the average time for completing each task are shortened approximately by half using PieceStack (with T.4 being an exception, improving from 109.384s to 30.458s), the performance improvement varied more:

for easy tasks like T.1, most users provided correct answers with both systems, and the accuracy is only increased from 99.58% to 100.00%, while it raised from 85.83% to 99.17% for the more complicated T.3. Fig. 1 shows more concrete comparisons.

Several subjective measures for PieceStack were also presented, with respect to the overall effectiveness, and the usefulness of some PieceStack designs involved in completing the tasks. All the measures were rated on a five-point scale, with 5 being strongly positive, and 1 strongly negative. Users' responses showed that they found Benchmark slightly more understandable (rated 4.35 versus 4.25 on average, with SD both being 0.75), and PieceStack easier to use (mean $= 4.1$, SD $= 0.79$ versus mean $= 4$, SD $= 1.17$). The significant rate difference for the usefulness problem conveyed that they agreed PieceStack to be much more useful (mean $= 4.7$, SD $= 0.47$ versus mean $= 2.9$, SD $= 0.72$). In addition, all encodings in PieceStack were positively rated, with clustering (mean $= 4.7$, SD $= 0.47$) and decomposition (mean$= 4.7$, SD $= 0.57$) being the most preferred ones and brushing (mean $= 4.25$, SD $= 1.02$) the least. Glyphs was rated 4.5 on average with SD $= 0.61$.

To further understand what caused these statistical and subjective differences, we analyzed the approaches users chose to solve the tasks. For the benchmark system, users' exploration processes were mostly identical. They would repeatedly reorder and stack the others on the layer of interest to compare it with the aggregation. They reported that this process was the only way to eliminate the visual illusion caused by baseline distortion. Meanwhile, the usage patterns of PieceStack varied more due to the extra cognitive load caused by the richer visual encodings. For example, only 12 users managed to take advantage of the glyphs (supposedly providing enough information) in T.1 and T.2, and 17 in T.3 and T.4. Others either chose to directly observe the cluster height at each time point (one for T.1 and T.2), or decomposed a whole cluster to compare its shape with the dashed line. In their self-reflections, three users felt that though admittedly the stacked graph was not expressive enough for the tasks, the information in PieceStack was, by contrast, overwhelming. They would need more time to get familiar with the complicated encodings and sometimes would still go directly with the observation or decomposition. Only when encountering more difficult tasks T.3 and T.4, they would try to consider the best approach possible and to turn to the glyphs. For the rest who did not use glyphs in the whole process, they reported that they were greatly attracted by the clustering results, which alleviated the baseline distortion and reduced the number of layers they must consider, and therefore ignored the glyphs. Despite of the complexity, all the users found some approaches, even if not necessarily the most ideal ones, that could make the comparisons among layers and aggregations more intuitive and less time consuming with PieceStack.

Overall, the users preferred our system since it provided an overview that could brief them with the general characteristics of stacked graphs and implemented interactions that allow them to target some parts with potential patterns. For instance, they agreed that the glyphs and cluster colors quantified the comparisons clearly. By contrast, the classical system only left them to explore blindly. The users further made some valuable points regarding of the clustering results' correctness and the intuitiveness of PieceStack. Four users felt when the number of layers reached the scale of hundreds or thousands, stacking them together would compress the layer so greatly that no information left except the aggregated values. In that sense, grouping layers into a number of clusters reasonable enough for display would require a coarser granularity process, and the similarity of layers in the

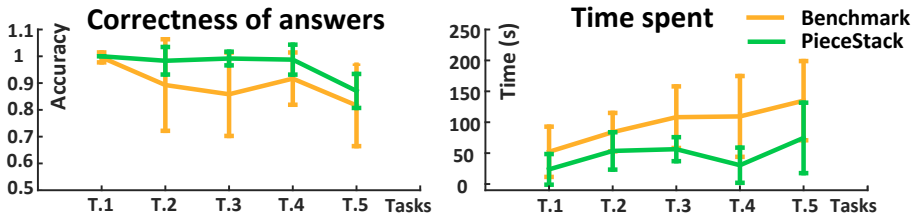

Figure 1: Error bar for the comparisons between the accuracy and efficiency of users completing tasks with two systems. The p-value for their accuracy differences are 0.165, 0.010, $1.621 \cdot 10^{-4}$, 0.001, 0.066 for T.1 to T.5 respectively, and the ones for the time spent are $3.135 \cdot 10^{-4}$, $9.135 \cdot 10^{-6}$, $3.003 \cdot 10^{-6}$, $1.203 \cdot 10^{-5}$, $1.568 \cdot 10^{-5}$.

same cluster could be questionable. Therefore, though the initial clustering result provides a general guideline, to deliver reliable results, they recommended to let the users decide the time intervals for more precise clustering. Besides, three users mentioned that for visual simplicity, encodings delivered on demand would be a better idea. For instance, the glyphs should only be displayed when the users want to check the contribution for some clusters in some intervals.

Appendix: Statistical summary of user study

Numerical results

Gender	Age	Degree	comp familiarity	Correctness and Timing																	
				T1				T2				T3				T4				T5			
				Benchmanrk		PieceStack		Benchmanrk		PieceStack		Benchmanrk		PieceStack		Benchmanrk		PieceStack		Benchmanrk		PieceStack	
				rate	time	rate	time (s)	rate	time (s)	rate	time (s)	rate	time (s)	rate	time (s)	rate	time (s)	rate	time (s)	rate	time (s)	rate	time (s)
M	24	PhD can	5	100.00%	49.15	100.00%	13.03	50.00%	119.13	100.00%	67.03	83.30%	96.39	100.00%	46.53	100.00%	28.49	100.00%	9.92	83.30%	55.49	83.30%	26.93
M	29	PhD can	5	100.00%	21.81	100.00%	15.40	85.30%	59.87	100.00%	44.00	83.30%	31.51	100.00%	51.39	91.67%	55.62	100.00%	23.07	75.00%	32.41	83.30%	16.55
M	22	UG Yr4	5	100.00%	49.23	100.00%	28.73	91.67%	89.60	100.00%	70.60	100.00%	78.23	100.00%	45.31	100.00%	81.27	100.00%	15.30	75.00%	103.15	91.67%	81.90
M	25	PhD can	5	100.00%	83.15	100.00%	31.82	100.00%	49.00	100.00%	30.66	100.00%	83.73	100.00%	36.85	100.00%	119.01	100.00%	41.88	83.33%	126.86	91.67%	120.61
M	22	UG Yr3	4	100.00%	90.72	100.00%	5.36	100.00%	58.39	100.00%	52.00	83.33%	113.22	100.00%	58.49	100.00%	42.42	100.00%	34.88	91.67%	146.48	83.33%	85.87
M	22	UR Yr3	4	100.00%	95.01	100.00%	60.64	100.00%	103.18	100.00%	101.47	83.33%	173.80	100.00%	72.01	100.00%	144.18	100.00%	20.59	100.00%	161.84	83.33%	44.05
F	21	UR Yr3	4	100.00%	39.78	100.00%	3.88	100.00%	89.04	100.00%	28.11	83.33%	73.82	100.00%	38.98	83.33%	11.32	100.00%	17.62	83.33%	22.05	83.33%	11.89
M	23	M.E.	5	100.00%	7.50	100.00%	5.36	66.67%	61.89	100.00%	25.61	75.00%	117.76	100.00%	49.20	75.00%	214.48	100.00%	108.58	75.00%	206.23	83.33%	58.44
F	24	M.S.	4	91.67%	53.27	100.00%	28.77	91.67%	62.63	83.33%	27.00	75.00%	153.47	100.00%	80.64	83.33%	117.70	100.00%	70.11	75.00%	202.54	83.33%	176.23
F	25	M.S.	4	100.00%	95.01	100.00%	39.78	50.00%	153.04	83.33%	133.11	50.00%	162.33	91.67%	90.64	91.67%	47.70	100.00%	30.11	75.00%	79.24	83.33%	16.23
M	23	PhD can	5	100.00%	49.89	100.00%	70.36	83.33%	106.11	100.00%	45.96	50.00%	141.51	91.67%	45.12	91.67%	235.73	100.00%	9.96	83.33%	174.05	83.33%	6.00
M	22	M.S.	4	100.00%	1.51	100.00%	0.51	100.00%	83.29	100.00%	73.08	100.00%	111.08	100.00%	44.75	75.00%	122.62	75.00%	25.81	75.00%	240.90	100.00%	161.73
F	23	M.S.	3	100.00%	170.93	100.00%	67.30	100.00%	100.86	100.00%	40.72	100.00%	217.20	100.00%	74.63	100.00%	166.09	100.00%	6.23	100.00%	153.52	100.00%	186.58
F	24	M.S.	4	100.00%	72.75	100.00%	69.31	100.00%	120.09	100.00%	38.06	91.67%	169.95	100.00%	94.80	83.33%	212.49	100.00%	4.60	91.67%	247.67	83.33%	94.16
M	22	UG Yr4	4	100.00%	18.30	100.00%	5.46	100.00%	125.11	100.00%	99.49	83.33%	114.32	100.00%	80.02	83.33%	163.51	100.00%	63.28	33.33%	103.66	83.33%	50.07
M	27	PhD can	5	100.00%	14.32	100.00%	1.76	100.00%	83.82	100.00%	43.72	100.00%	94.34	100.00%	46.23	75.00%	73.56	100.00%	4.71	100.00%	141.68	83.33%	23.09
M	22	UG Yr4	5	100.00%	56.34	100.00%	3.44	100.00%	47.24	100.00%	31.01	100.00%	76.84	100.00%	58.88	100.00%	33.19	100.00%	11.72	100.00%	84.27	100.00%	99.39
M	23	M.S.	4	100.00%	11.69	100.00%	4.45	100.00%	56.72	100.00%	69.27	100.00%	80.49	100.00%	53.39	100.00%	86.96	100.00%	11.72	83.33%	86.86	83.33%	99.39
F	22	UG Yr3	4	100.00%	28.78	100.00%	10.30	66.67%	32.83	100.00%	18.07	75.00%	36.53	100.00%	31.48	100.00%	113.01	100.00%	78.73	66.67%	176.62	91.67%	116.98
F	24	PhD can	5	100.00%	34.47	100.00%	9.22	100.00%	77.21	100.00%	32.33	100.00%	36.53	100.00%	28.28	100.00%	118.32	100.00%	20.33	83.33%	151.78	83.33%	16.55
p-value				rate		0.1649384		rate		0.009869771		rate		0.000162195		rate		0.001227612		rate		0.065739396	
				time		0.000213509		time		9.13489E-06		time		3.00323E-06		time		1.20332E-05		time		0.000156845	
MEAN	23.450			0.996	52.181	1.000	23.744	0.893	83.953	0.983	53.565	0.858	108.153	0.992	56.381	0.917	109.384	0.988	30.458	0.817	134.865	0.871	74.632
SD	1.932			0.019	40.653	0.000	24.782	0.171	31.086	0.051	30.231	0.156	49.823	0.026	19.398	0.097	65.352	0.056	28.439	0.152	64.068	0.063	56.973
MAX	29.000			1.000	170.930	1.000	70.360	1.000	153.040	1.000	133.110	1.000	217.200	1.000	94.800	1.000	235.730	1.000	108.580	1.000	247.670	1.000	186.580
MIN	21.000			0.917	1.510	1.000	0.510	0.500	32.830	0.833	18.070	0.500	31.510	0.917	28.280	0.750	11.320	0.750	4.600	0.333	22.050	0.833	6.000
MEDIAN	23.000			1.000	49.190	1.000	11.665	1.000	83.555	1.000	43.860	0.833	103.735	1.000	50.295	0.958	115.355	1.000	20.460	0.833	144.080	0.833	70.170
MODE	22.000			1.000	95.010	1.000	5.360	1.000	N/A	1.000	N/A	1.000	36.530	1.000	N/A	1.000	N/A	1.000	11.720	0.750	N/A	0.833	16.550

Questionnare

Questions			1	2	3	4	5	Remarks	MEAN	STD	MEDIAN	MIN	MAX	MODE
The system is understandable?	Benchmark				3	7	10		4.35	0.745	4.5	3	5	5
	PieceStack				3	9	8		4.25	0.716	4	3	5	4
The system is easy to use?	Benchmark		2		1	10	7		4	1.170	4	1	5	4
	PieceStack				4	10	6		4.1	0.718	4	3	5	4
The system is useful overall?	Benchmark			7	8	5			2.9	0.788	3	2	4	3
	PieceStack					6	14		4.7	0.470	5	4	5	5
Encodings and interactions are useful?	Benchmark	Layers		1	6	9	4		3.8	0.834	4	2	5	4
		Reordering		1	6	9	4		3.8	0.834	4	2	5	4
		Baseline Straighten	5	2	7	3	3		2.85	1.387	3	1	5	3
	PieceStack	Clusters				6	14		4.7	0.470	5	4	5	5
		Glyphs			1	8	11		4.5	0.607	5	3	5	5
		Brushing		2	2	5	11		4.25	1.020	5	2	5	5
		Decomposition			2	5	13		4.7	0.571	5	3	5	5

Comments
Both system
- Coloring: being easily attracted by the dark colors (e.g., dark blue) and will choose to test those ones first.
- Too many layers are stacked together + generate too many colors.
Intuitiveness and Information on Demand
- Too many information encoded. Would be better to display on demand (e.g., user decide which interval's glyph should be displayed).
- Outlier's glyph should also be added.
- now only the cluster information is displayed. When brushing a layer, better hide all other glyphs, and show only the selected layer's contribution, so the layers need not to be dragged out individually.
Correctness
- If the cluster's contribution could really be generated to all layers involved when clustered with coarser granularity process.
- Should let users to define the cluster period.
Decomposition
- Should always keep an unbroken graph all the time, and generate new graphs whenever decompsiting.
- Should add the included layers in the thumbnail view.