

Querying Billion-Edge Evolving Property Graphs with Portal

Julia Stoyanovich New University, USA

stoyanovich@nyu.edu @stoyanoj

portaIDB.github.io



Is there a path from Ann to Cat?





Is there a path from Ann to Cat?



point-based semantics



TGraph: evolving property graph

TGraph: evolving property graph



▶ Definition 1. TGraph is a six-tuple $\mathcal{G} = (V, E, L, \rho, \xi^T, \lambda^T)$, where

- V is a finite set of nodes (or vertices), E is a finite set of edges, $V \cap E = \emptyset$;
- \blacksquare L is a finite set of property labels;
- $\rho: E \to (V \times V)$ is a total function that maps an edge to its source and destination nodes;
- $\xi^T : (V \cup E) \times \Omega^T \to B$ is a total function that maps a node or an edge, and time point, to a Boolean, indicating existence of that node or edge; and
- $\lambda^T : (V \cup E) \times L \times \Omega^T \to val$ is a partial function that maps a node or an edge, a property label, and a time point, to a property value.

[Moffitt, Stoyanovich, DBPL 2017]



TGraph: evolving property graph





[Moffitt, Stoyanovich, DBPL 2017]



TGraph algebra (TGA)





• Operators

- temporal variants of standard graph operators: union, intersection, difference, slice, subgraph, filter, Pregel-style analytics
- novel operator: temporal window-based zoom
- TGA is **compositional**
- Operations maintain model integrity under point-based semantics

[Moffitt, Stoyanovich, **DBPL 2017**]



Which institutions collaborate?







Which institutions collaborate?



temporal attribute-based zoom



snapshot reducibility

Who is Bob's BFF?





January 2019

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
30	31	1	2	3	4	5
		New Year's Day				
6	7	8	9	10	11	12
12	14	15	16	17	10	10
13	14	15	10	17	18	15
20	21	22	23	24	25	26
	Martin Luther King Day					
27	28	29	30	31	1	2
L						Pade associated from tell codition decomposite

© Calendarpedia® www.calendarpedia.com

2017	20	18	201	19
Jernany 1 2 3 4 5 6 7 1 9 1 <th1< th=""></th1<>	Name Jernmany Jernmany 1 3 4 5 6 1 3 4 5 6 10 1 2 3 4 5 6 6 10	February Ise Ma Ta We Ta Fe Ise 1 2 3 5 6 7 8 9 46 16 4 7 8 0 46 10 46 17 <t< th=""><th>January Se Met 7s Wet 7s Fe Fe Met 7s Se 7s 7s</th><th>February 1 2 7 7 7 8 8 7 1 8 1<</th></t<>	January Se Met 7s Wet 7s Fe Fe Met 7s Se 7s	February 1 2 7 7 7 8 8 7 1 8 1<
Manageh Ag Se 1 Mail No. 1 Mail	March March <th< td=""><td>April 9 10 10 10 10 10 10 10 10 10 10 10 10 10</td><td>March 34 March 76 Wei 76 76 76 76 5 4 5 4 7 8 8 1 5 4 5 4 7 8 1 5 4 5 4 7 8 1 6 10 20 21 22 20 1 6 20 21 22 20 1 7 10 10 20 21 22 20 1 7 10 10 20 10 10 10 1 7 10 10 20 10 10 10 1 7 10 10 10 10 10 1 7 10 10 10 10 1 7 10 10 10 10 1 7 10 10 10 1 7 10 10 10 1 7 10 10 10 1 7 10 1 7</td><td>April 10 10 10 10 10 10 10 10 1 2 3 4 5 1 3 5 1 5 10 1 4 5 10 10 10 10 10 1 2 2 20 24 25 28 27 1 2 10 10</td></th<>	April 9 10 10 10 10 10 10 10 10 10 10 10 10 10	March 34 March 76 Wei 76 76 76 76 5 4 5 4 7 8 8 1 5 4 5 4 7 8 1 5 4 5 4 7 8 1 6 10 20 21 22 20 1 6 20 21 22 20 1 7 10 10 20 21 22 20 1 7 10 10 20 10 10 10 1 7 10 10 20 10 10 10 1 7 10 10 10 10 10 1 7 10 10 10 10 1 7 10 10 10 10 1 7 10 10 10 1 7 10 10 10 1 7 10 10 10 1 7	April 10 10 10 10 10 10 10 10 1 2 3 4 5 1 3 5 1 5 10 1 4 5 10 10 10 10 10 1 2 2 20 24 25 28 27 1 2 10 10
Mage Mage <th< td=""><td>Open Max Max<td>Jump Jump <th< td=""><td>Mary Mary <th< td=""><td>June June <thjune< th=""> June June <thj< td=""></thj<></thjune<></td></th<></td></th<></td></td></th<>	Open Max Max <td>Jump Jump <th< td=""><td>Mary Mary <th< td=""><td>June June <thjune< th=""> June June <thj< td=""></thj<></thjune<></td></th<></td></th<></td>	Jump Jump <th< td=""><td>Mary Mary <th< td=""><td>June June <thjune< th=""> June June <thj< td=""></thj<></thjune<></td></th<></td></th<>	Mary Mary <th< td=""><td>June June <thjune< th=""> June June <thj< td=""></thj<></thjune<></td></th<>	June June <thjune< th=""> June June <thj< td=""></thj<></thjune<>
Joing Analy Analy Analy a bits b	Specify July	Analysis Bar Ma To No No <td>Johng In <thi< td=""><td>August a bio Ta via Ta Fr Ba 4 5 6 7 8 9 9 14 10 13 14 16 16 9 14 19 20 17 28 29 30 31 28 27 28 29 30 31</td></thi<></td>	Johng In In <thi< td=""><td>August a bio Ta via Ta Fr Ba 4 5 6 7 8 9 9 14 10 13 14 16 16 9 14 19 20 17 28 29 30 31 28 27 28 29 30 31</td></thi<>	August a bio Ta via Ta Fr Ba 4 5 6 7 8 9 9 14 10 13 14 16 16 9 14 19 20 17 28 29 30 31 28 27 28 29 30 31
September Octo September International Internation International International </td <td>Object Seglective 0 10</td> <td>October Im Mo To Mo To Mo Mo<td>Sectors Sectors <t< td=""><td>October He He 171 We 175 He He 1 2 3 4 He 174 He 4 7 8 9 10 11 02 13 94 15 14 17 18 9 14 21 22 23 24 25 M 17 28 29 20 21</td></t<></td></td>	Object Seglective 0 10	October Im Mo To Mo To Mo Mo <td>Sectors Sectors <t< td=""><td>October He He 171 We 175 He He 1 2 3 4 He 174 He 4 7 8 9 10 11 02 13 94 15 14 17 18 9 14 21 22 23 24 25 M 17 28 29 20 21</td></t<></td>	Sectors Sectors <t< td=""><td>October He He 171 We 175 He He 1 2 3 4 He 174 He 4 7 8 9 10 11 02 13 94 15 14 17 18 9 14 21 22 23 24 25 M 17 28 29 20 21</td></t<>	October He He 171 We 175 He He 1 2 3 4 He 174 He 4 7 8 9 10 11 02 13 94 15 14 17 18 9 14 21 22 23 24 25 M 17 28 29 20 21
Nonember No <	United in the second	December Ma Ta We Ta Fe Ne 3 4 5 4 7 1 3 4 5 4 7 1 4 1 12 13 4 14 46 17 18 19 20 21 21 19 24 15 26 27 21 21 21 29 21 20 27 28 29 21 21	November 1 36 Mi 75 Wei 75 F 10 1 1 3 4 5 6 7 8 1 9 11 213 Mi 15 10 1 1 16 11 213 Mi 15 10 10 10 17 18 10 20 21 22 22 23 18 20 20 21 20 22 23 24	December Fit Mite Tit Vite Tit Fit Mite 1 2 5 4 5 6 7 8 9 5 1 12 3 44 8 11 12 13 44 5 6 7 86 16 17 145 19 20 17 3 82 233 24 28 28 28 27 38 93 30 34 5 5 7 38
Federal Holidays 2017 Sea 1 New Yarki Dig May Sea 1 New Yarki Dig Marki Met Sea 1 New Yarki Dig Marki Met Sea 1 New Yarki Met	Construction Federal Holidays 2018 w Jan 1 New Yark Usy a0 and the second	Caleriviaryandia Data Caleria Day No 10 Vatros Day No 10 Vatros Day No 10 Vatros (No 1000) No 10 Vatros (No 1000) No 10 Vatros (No 1000) No 20 Thekaping Day Da 20 Ovideo Day	Federal Holidays 2019 Jac 1 Ner Yerk Day Jac 9 Nerb Later Eng Day He 27 Nerbert Day Net 7 Nerbert Day Jac 8 Internet Day	Calenciaryon & Science & Series Science & Series Science & Series Science & Series Science & Series Science & Series Science & Series



temporal window-based zoom

window	points	interval
Q1	1, 2, 3	[1,4)
Q2	4, 5, 6	[4,7)
Q3	7, 8, 9	[7,10)

friend in *some* snapshot during a given quarter

window width=3, nodes=EXIST, edges=EXIST, node values=first, edge values=any





temporal window-based zoom

window	points	interval
Q1	1, 2, 3	[1,4)
Q2	4, 5, 6	[4,7)
Q3	7, 8, 9	[7,10)

friend in *every* snapshot during a given quarter

window width=3, nodes=ALL, edges=ALL, node values=last, edge values=any



extended snapshot reducibility

Where is my bus?





Most frequent taxi route in NYC: Penn Station to Grand Central

Why?

NYC TLC data 07/2015 - 06/ 2016

Where is my bus?





Midtown to airports:

LGA is much smaller than JFK, but gets far more taxi traffic.

Why?

NYC TLC data 07/2015 - 06/ 2016

Where is my bus?





- data: pick-up / drop-off time & location, fare, passenger count
- trips represented as a TGraph
 - nodes represent locations, with latitude / longitude coordinates as an attribute; a node exists from the time of the first incoming or outgoing trip until the time of the last trip
 - edges represent trips, with duration, fare etc as attributes; an edge exists for the duration of the trip
 NY(

NYC TLC data 07/2015 - 06/ 2016



Portal: implementation

 $\blacksquare Graph X$



goal: principled and systematics support for usable, scalable and extensible analysis of evolving graphs





















"Representative Graphs" (snapshot sequence)







"Vertex Edge" (nested relational)

ID	Attributes	Т
Ann	Type=person, School=NYU	[1,7)
Bob	Type=person	[2,5)
Bob	Type=person, School=CMU	[5 <i>,</i> 9)
Cat	Type=person, School=NYU	[1,9)

Vertices

|--|

ID	V1	V2	Attributes	Т
e1	Ann	Bob	Type=co-author	[2,7]
e2	Bob	Cat	Type=co-author	[7,9)





"One Graph" (GraphX graph)





Performance highlights

Datasets



WikiTalk	SNB(LDBC Social Network Benchmark)	nGrams			
Communication graph	Friendship graph	Word co-occurrence graph			
# of nodes: 2.9 M	# of nodes: 3.3 M	# of nodes: 48 M			
# of edges: 10.7M	# of edges: 202 M	# of edges: 1.32B			
# of intervals: 179	# of intervals: 36	# of intervals: 328			



Cluster : 16-workers in-house cluster Workers: 4 cores and 32 GB of RAM

Zoom on nGrams







Wrapping up

Ongoing work



• TGA / Portal

- Declarative language, query optimization
- Data generation, benchmarking
- Applications: socioeconomic studies
- Journeys, temporal regular path queries: semantics, complexity of evaluation, implementation

Take-aways



- TGraph: a logical model of property graphs with time
- TGA: a compositional temporal graph algebra under point semantics
- Portal: a library on top of Apache Spark, interoperable with SparkSQL and other libraries
- Performs well on billion-edge graphs with interesting evolution patterns
- NYC Taxi use case, working on others



In this example we compute the degree centrality of DBLP over time. Degree centrality is a simple measure the uniformity of influence in a graph. For more information, see the definition here.

The most centralized graph is a "star graph". A central node is connected to every other, and every outside node is connected only to the center.

The least centralized graph is a "circle graph", where all nodes have the same exact degree.











Vera Moffitt

Amir Aghasadeghi



Sebastian Schelter

[CAREER] Querying evolving graphs, 03/2018-





Thank you!



Back-up

aZoom ("One Graph")



Algorithm 3 aZoom^{*T*} over OG

Require: Skolem function $f_s: V \Rightarrow \mathbb{N}$; Aggregation function f_{aaa} : $V \times V \Rightarrow V$ 1: $V' \leftarrow V$.flatMap $\{v \Rightarrow$ $v.history.map\{(_, attr) \Rightarrow$ 2: $v.copyWithIdAndAttributes(f_s(v.vid), attr) \}$ 3: 4: .groupBy { $v \Rightarrow v.vid$ } 5: .reduce { $(v_a, v_b) = f_{aqq}(v_a, v_b)$ } 6: $E' \leftarrow E . \operatorname{map} \{ e \Rightarrow$ Ann $h \leftarrow$ recompute history(e) 7: type=person school=MIT e.copyWithVidsAndHistory($f_s(e.v1.vid)$), T = [1, 7) 8: $f_{s}(e.v2.vid), h)$ 9: return new TGraph G(V', E')



wZoom ("Vertex Edge")

				v	2	т			1	1		l			
		_		<u> </u>	ů	•		<u>v</u>	а	T	r_v				
⊻	а	Т		Bob	{ }	[1, 4]		Bob	[{}]	[1, 4)	0.66				1
Boh	11	234	Align with window	Boh	13	[4 7)	Group by		,	(-) ./	0.00	Eiltor by	<u>v</u>	а	Т
000	13	2,3,4	1, 2, 3 🗲 [1,4)		U	[4,7]			[{},			FILET Dy			
			4, 5, 6 → [4,7)	Bob	{ s=CMU }	[4,7)	intorval	Bob	{ s=CMU }]	[4, 7):	1.0	r_v	Dah		[4 7)
Bob	{ s=CMU }	5,6,7,	7 8 9 -> [7 10]				interval					Resolve:		{ S=CIVIU }	[4, 7]
		8	, , , , , , , , [7,10]	Bob	{ s=CMU }	[7, 10)		Bob	[{ s=CMU }]	[7, 10)	0.33	last			

wZoom ("Vertex Edge")



Algorithm 5 wZoom^Tover VE **Require:** resolve functions f_{v} , f_{e} ; quantifiers r_{v} , r_{e} ▶ *Computation of new intervals* 1: 2: $I' \leftarrow I. \operatorname{map} \{ i \Rightarrow (i, \operatorname{computeNewInterval}(i)) \}$ ▶ *Vertex aggregation for new intervals* 3: 4: $V' \leftarrow V.\mathbf{join}(I').\mathbf{on}\{(v, (i, n)) \Rightarrow v.n == i\}$ $.map \{ (v, (i, newInterval)) \Rightarrow$ 5: v.copyWithNewInterval(newInterval)} 6: $.groupBy{ v \Rightarrow (v.id, v.interval) }$ 7: .filter {(*i*, vertices) \Rightarrow match threshold(vertices, r_{τ})} 8: .reduceByKey{($(v_a), (v_b)$) $\Rightarrow f_v(v_a, v_b)$ } 9: ▶ Edge aggregation for new intervals 10: 11: $E' \leftarrow E.join(I').on\{(e, (i, n)) \Rightarrow e.interval == n\}$ $.map \{ (e, (i, newInterval)) \Rightarrow$ 12: *e*.copyWithNewInterval(*newInterval*)} 13: $.groupBy{e \Rightarrow (e.id, e.interval)}$ 14: .filter { $(i, edges) \Rightarrow$ match_threshold(edges, r_e)} 15: .reduceByKey{($(e_a), (e_b)$) $\Rightarrow f_e(e_a, e_b)$ } 16: if $r_{\upsilon} > r_e$ then ► Dangling edge removal 17: $E'' \leftarrow E'$.semijoin(V')18: $.on\{(e, v) \Rightarrow e.vid^{1} == v.id \text{ and in interval}(e, v)\}$ $E''' \leftarrow E''$.semijoin(V') 19: .on { $(e, v) \Rightarrow e.vid2 == v.id$ and in interval(e, v) } 20: return new TGraph (V', E''')