CoreFlow

Extracting and Visualizing Branching Patterns from Event Sequences

Zhicheng "Leo" Liu, Bernard Kerr, Mira Dontcheva

Justin Grover, Matthew Hoffman, Alan Wilson

Adobe

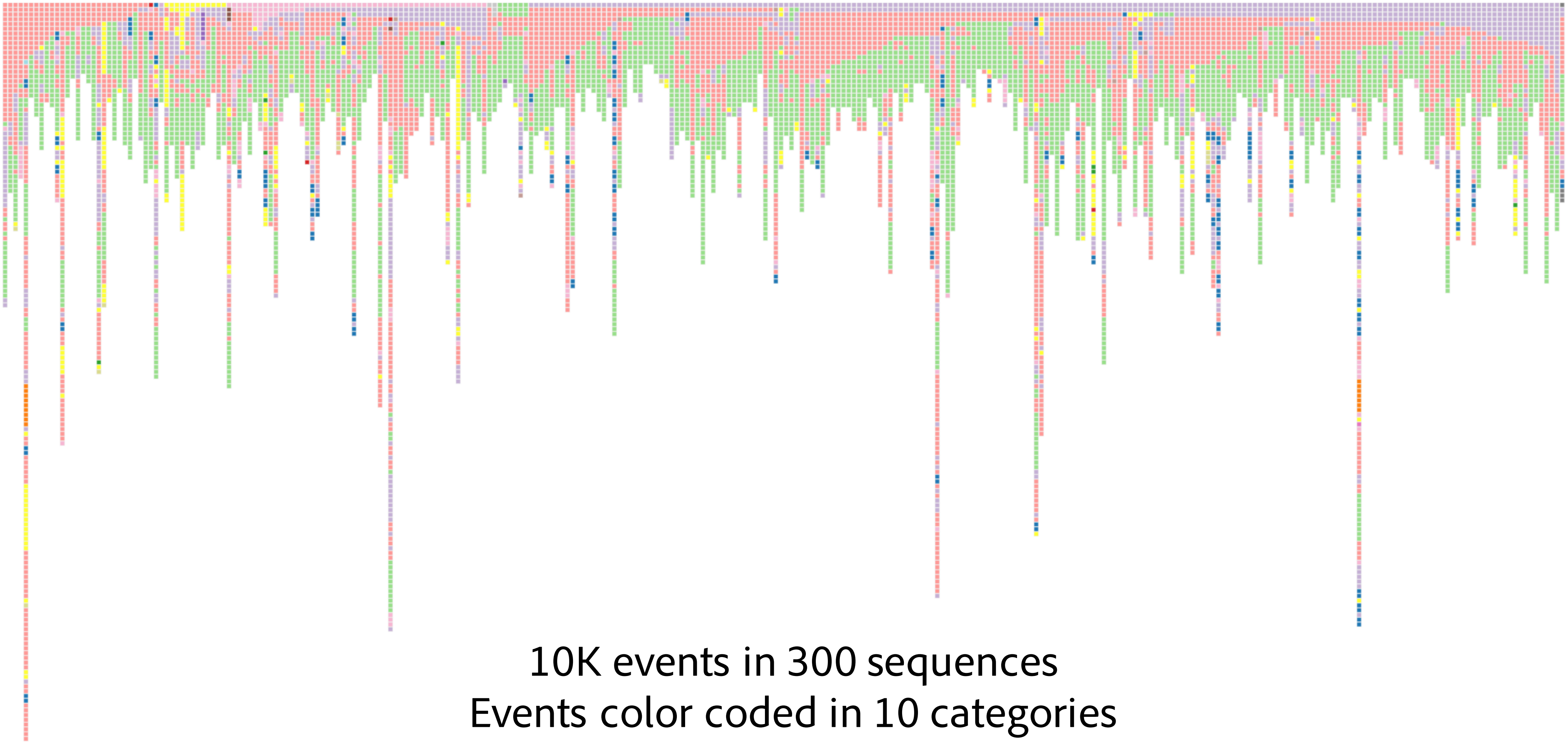
Event Sequences: Clickstreams

Timestamp	Page Click	OS
06/29/2016 16:01:20	adobe.com	OS X
06/29/2016 16:03:04	adobe.com/creativecloud/photography.html	OS X
06/29/2016 16:03:29	creative.adobe.com/products/download/ccpp	OS X
06/29/2016 16:05:12	creative.adobe.com:Authenticated	iOS
06/29/2016 16:06:23	creative.adobe.com:Photography:Join:1:AdobeIDForm:Page	iOS
06/29/2016 16:07:34	creative.adobe.com:Photography:Join:2:ReviewMembershipDetails:Page	iOS
06/29/2016 16:07:58	creative:AnywareCheckout:checkoutLoaded	iOS
06/29/2016 16:08:24	creative.adobe.com:Photography:Join:3:PaymentInfo:Page	iOS
06/29/2016 16:08:51	creative:AnywareCheckout:validateOrder	iOS
06/29/2016 16:09:06	creative.adobe.com:Join:Checkout:Order:Validated	iOS
06/29/2016 16:09:21	creative:AnywareCheckout:orderValidated	iOS
06/29/2016 16:11:32	creative.adobe.com:Photography:Join:4:ConfirmOrder:Page	iOS

Event Sequences: Application Logs

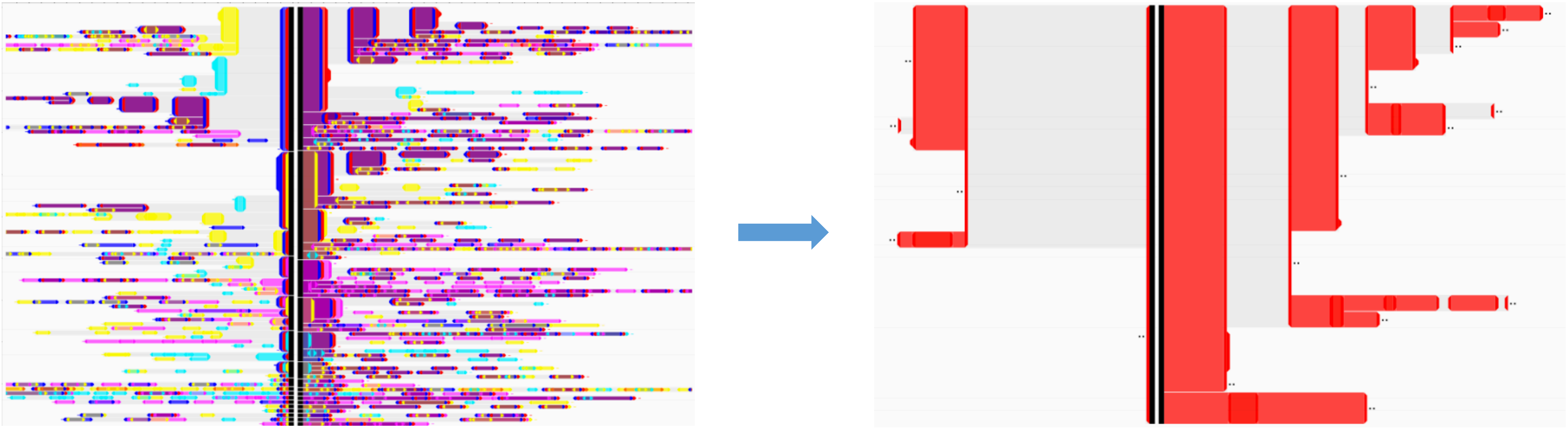
Timestamp	Operation
01/08/2017 08:01:12	editor: titleChanged
01/08/2017 08:01:32	projectCreated
01/08/2017 08:01:48	editor: titleChanged
01/08/2017 08:01:48	editor: titleChanged
01/08/2017 08:02:09	editor: subTitleChanged
01/08/2017 08:02:54	imageChosen
01/08/2017 08:03:17	editor: titleImageChanged
01/08/2017 08:03:17	editor: titleImageChanged
01/08/2017 08:03:55	imageChosen
01/08/2017 08:04:03	editor: titleImageChanged
01/08/2017 08:05:22	editor: textContentITemAdded

Challenges in visualizing event sequences



10K events in 300 sequences
Events color coded in 10 categories

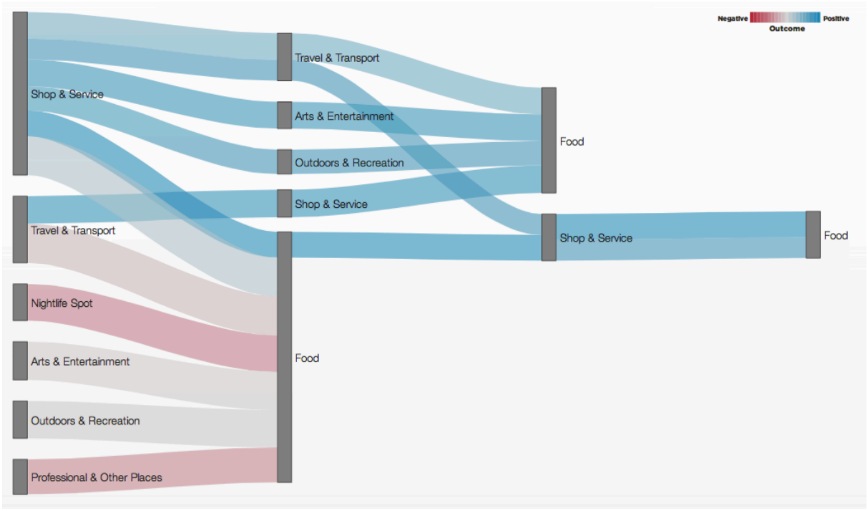
Human-in-the-loop Filtering and Aggregation



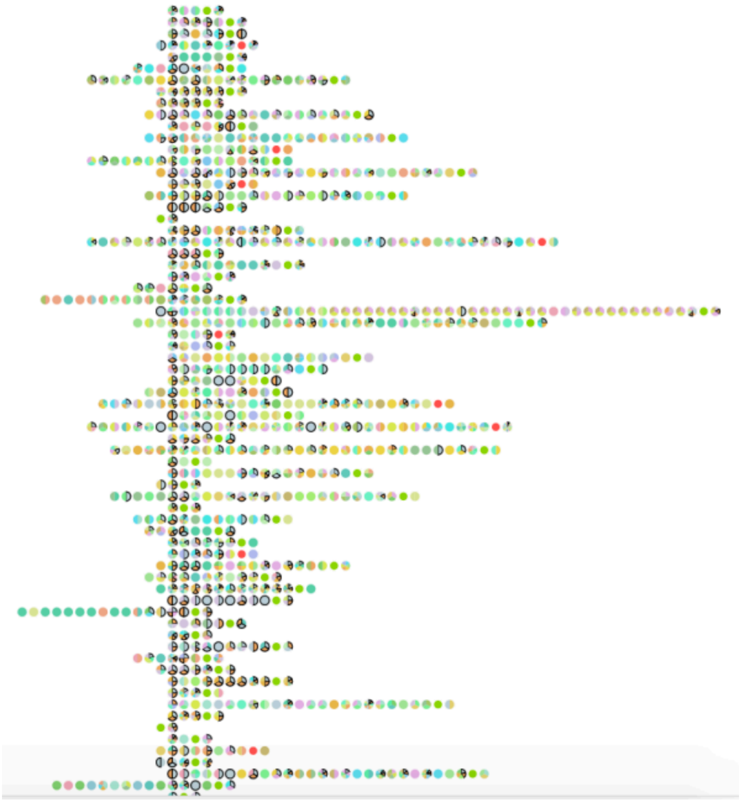
"Temporal Event Sequence Simplification", Monroe et. al. 2013

"Coping with Volume and Variety in Temporal Event Sequences: Strategies for Sharpening Analytic Focus", Du et. al. 2016

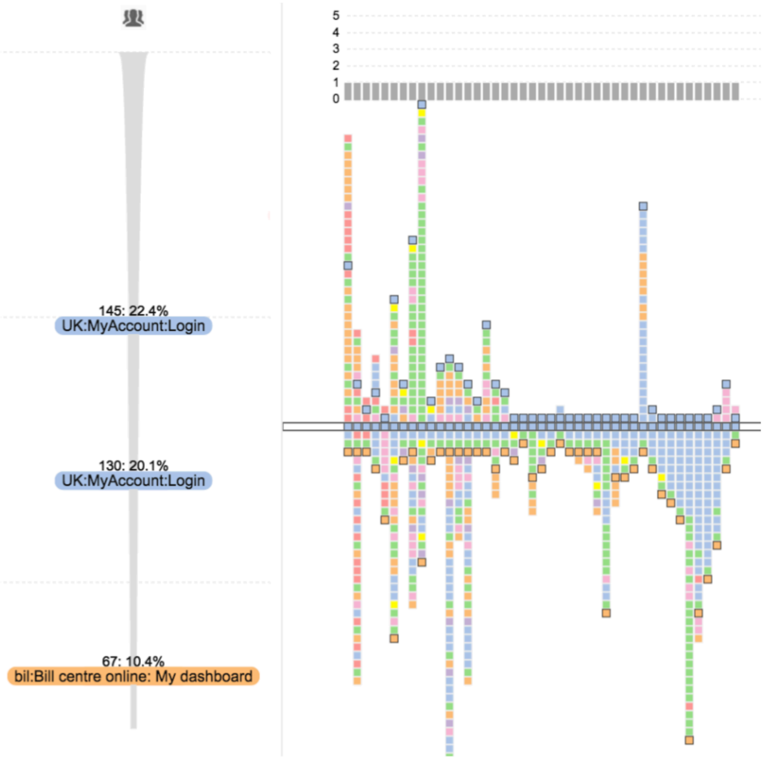
Automatic Extraction of Patterns



Frequence
Perer and Wang, 2014

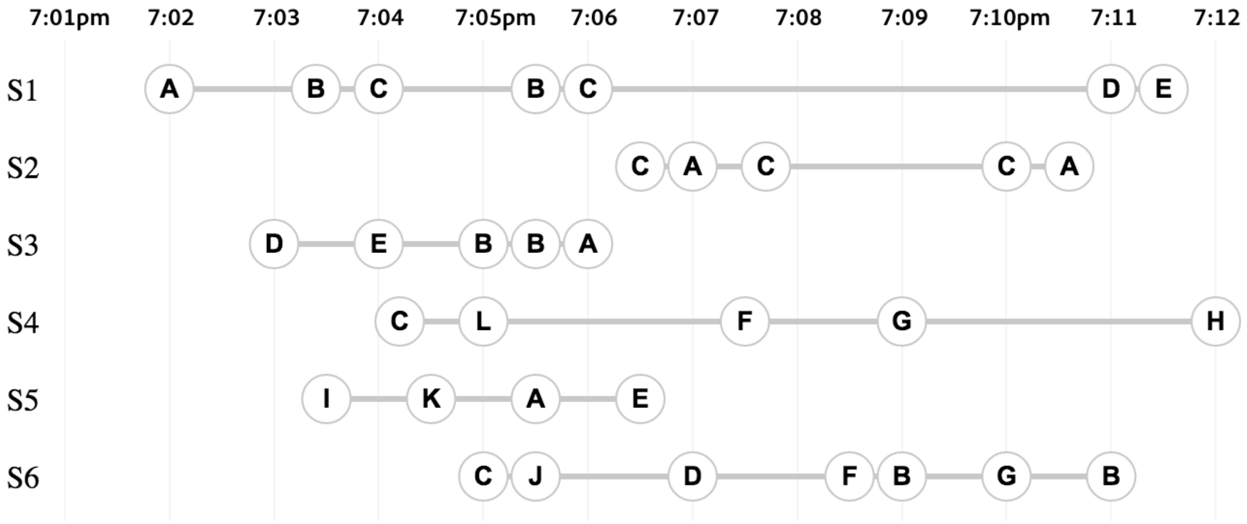


Peekquence
Kwon et. al. 2016

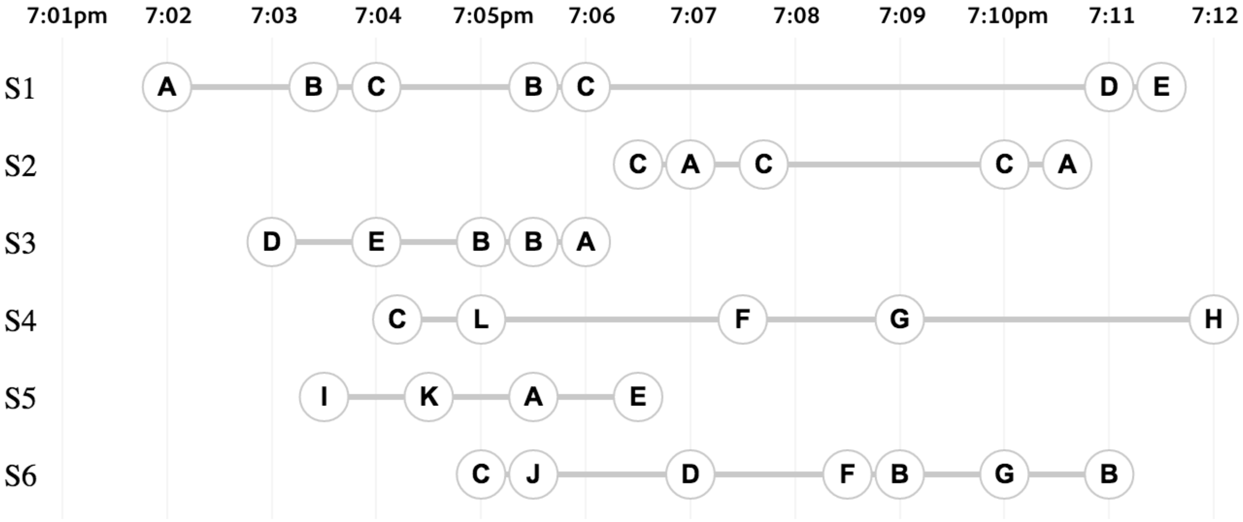


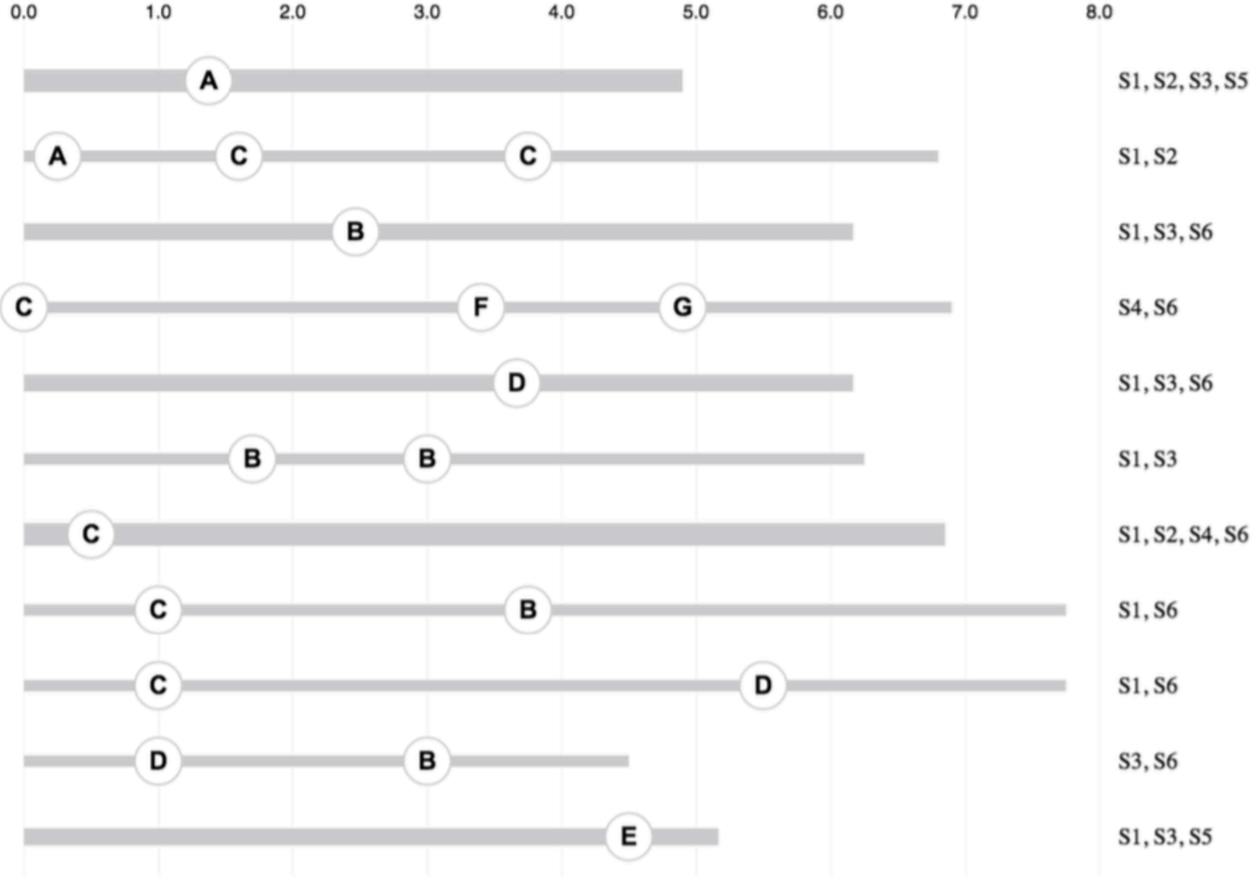
Patterns & Sequences
Liu et. al. 2016

Automatic Extraction of Sequential Patterns

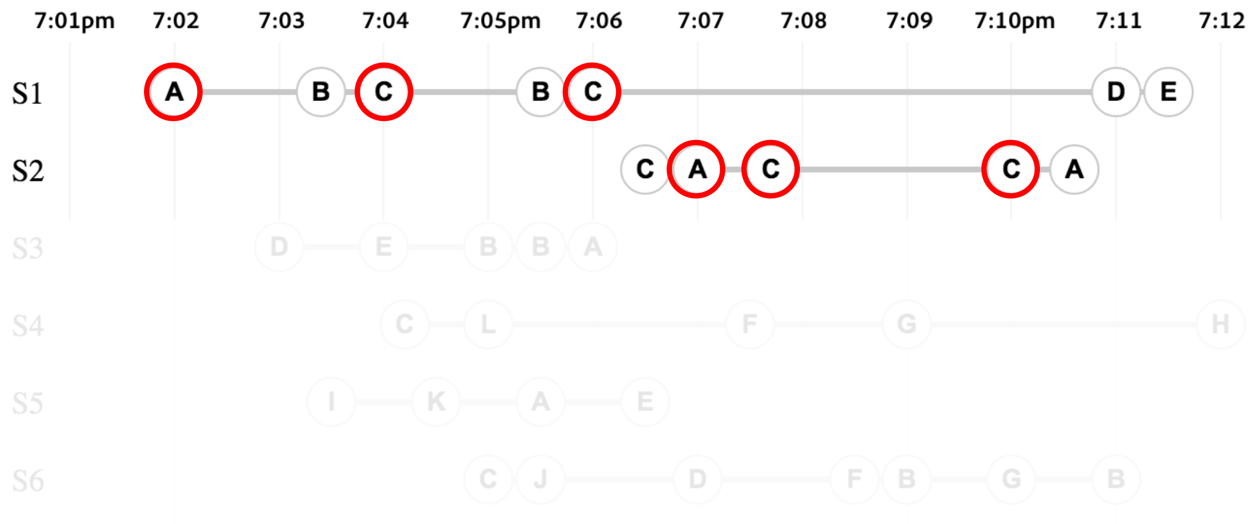


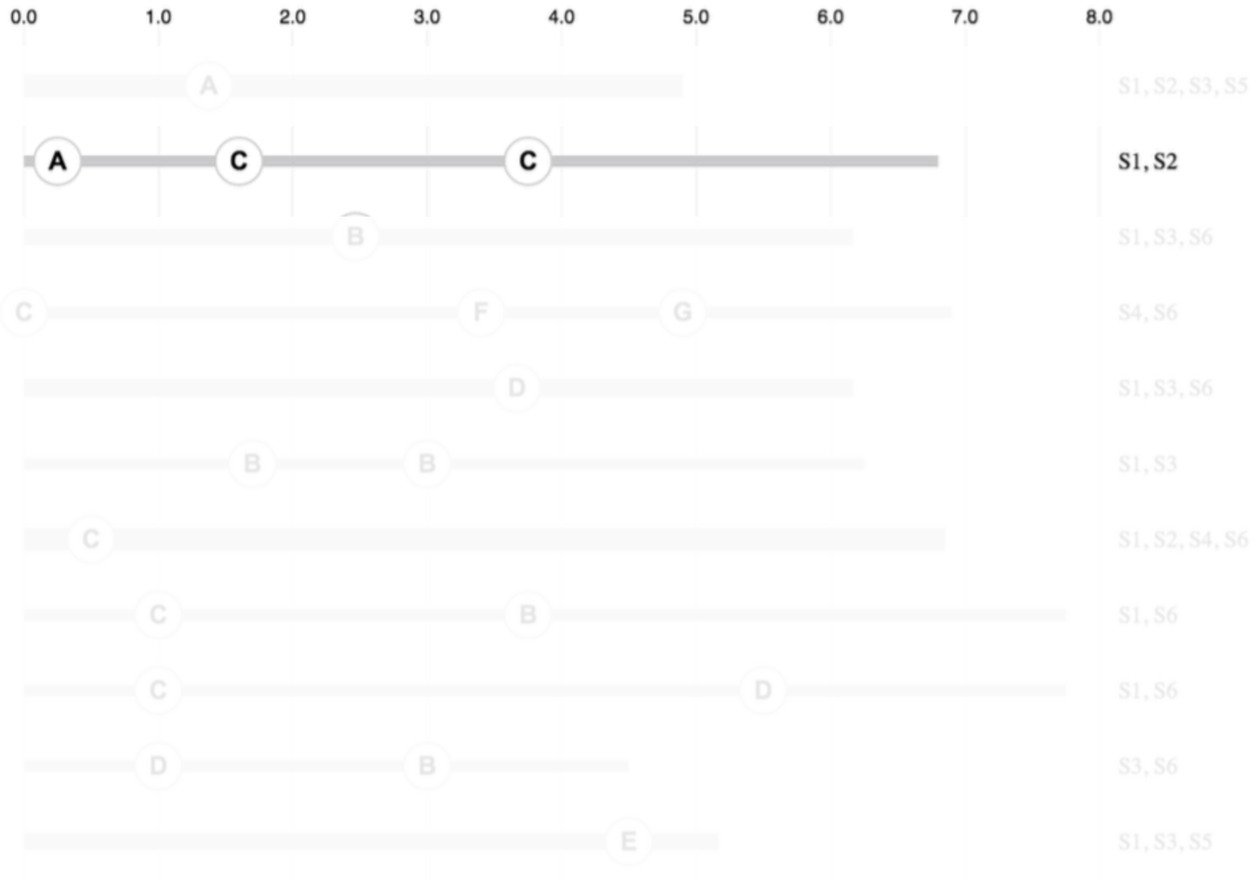
Automatic Extraction of Sequential Patterns





Automatic Extraction of Sequential Patterns





Potential Problems

Interpretability

relationship between patterns

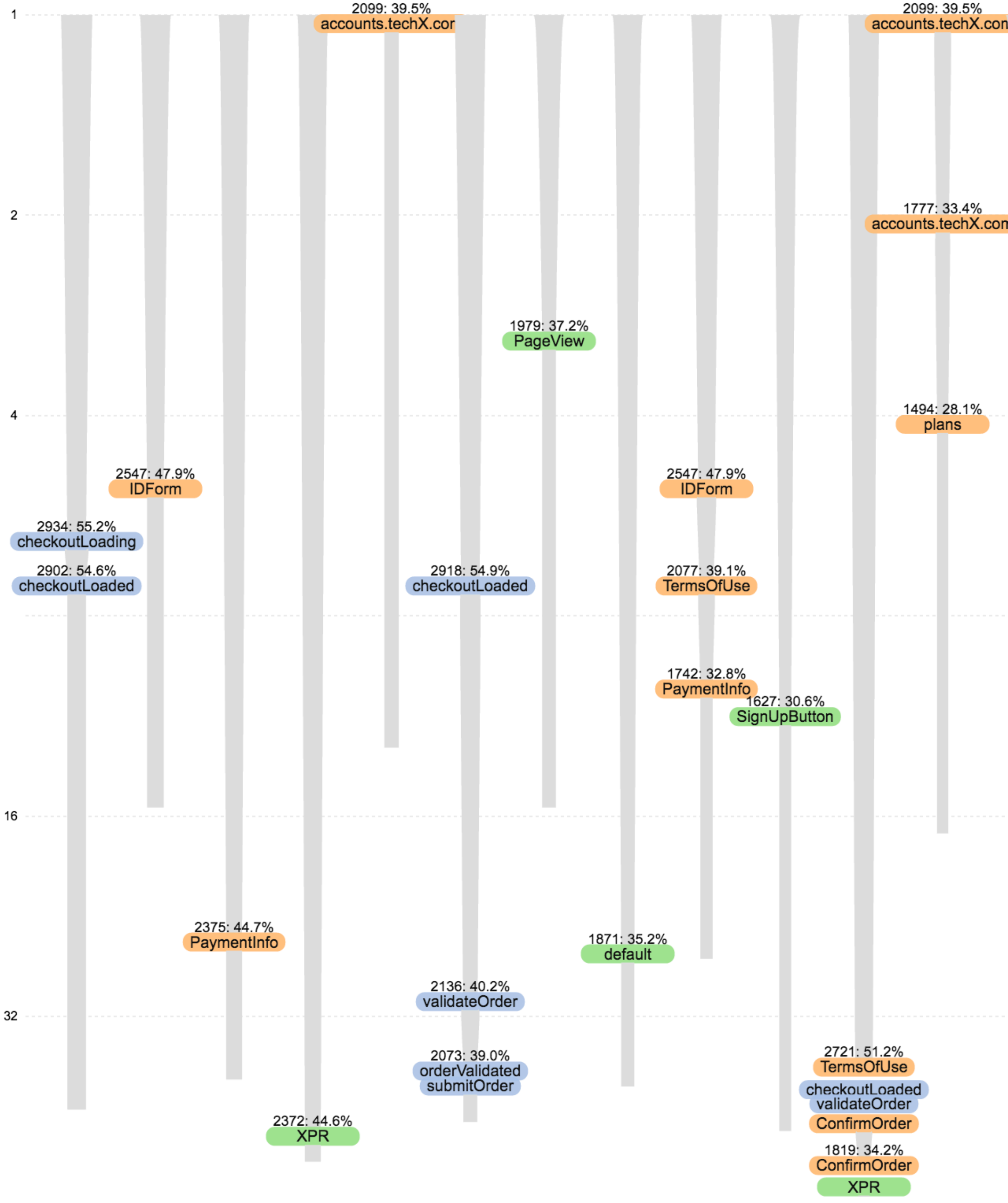
number of patterns to review

Scalability

computation cost

Utility

frequent = interesting/useful?



Potential Problems

Interpretability

relationship between patterns

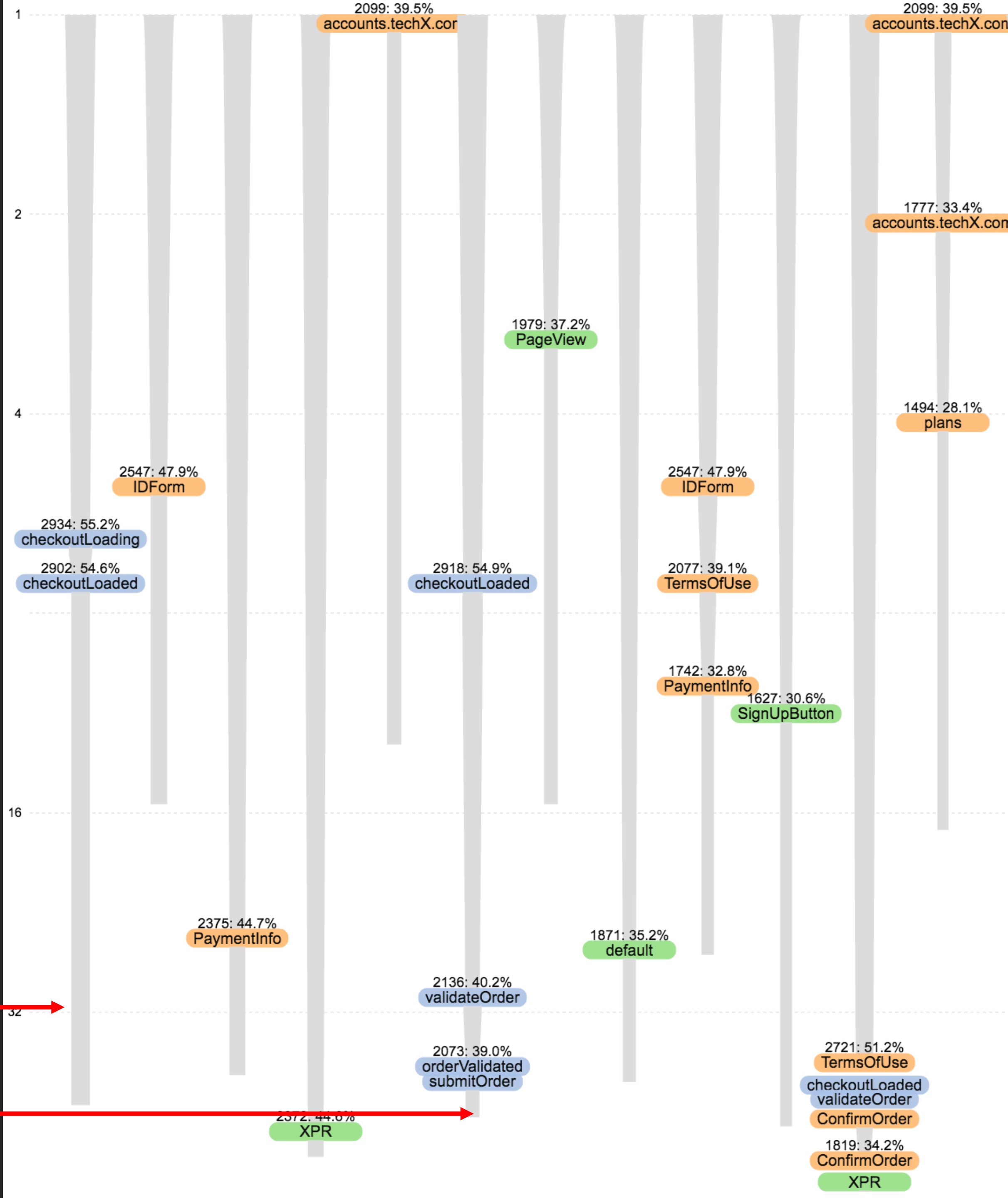
number of patterns to review

Scalability

computation cost

Utility

frequent ≠ interesting/useful?



checkoutLoading -> checkoutLoaded

checkoutLoaded -> validateOrder -> OrderValidated -> submitOrder

Potential Problems

Interpretability
relationship between patterns
number of patterns to review

	# of Input Sequences	# of Maximal Patterns
Dataset 1	2512	6079
Dataset 2	2712	6241
Dataset 3	2739	5130

Scalability
computation cost

Utility
frequent = interesting/useful?

Potential Problems

Interpretability
relationship between patterns
number of patterns to review

Scalability
computation cost

	# of Input Sequences	Computational Time
Dataset 1	2512	6.18 minutes
Dataset 2	2712	7.16 minutes
Dataset 3	2739	6.05 minutes

Utility
frequent = interesting/useful?

Potential Problems

Interpretability

relationship between patterns
number of patterns to review

Scalability

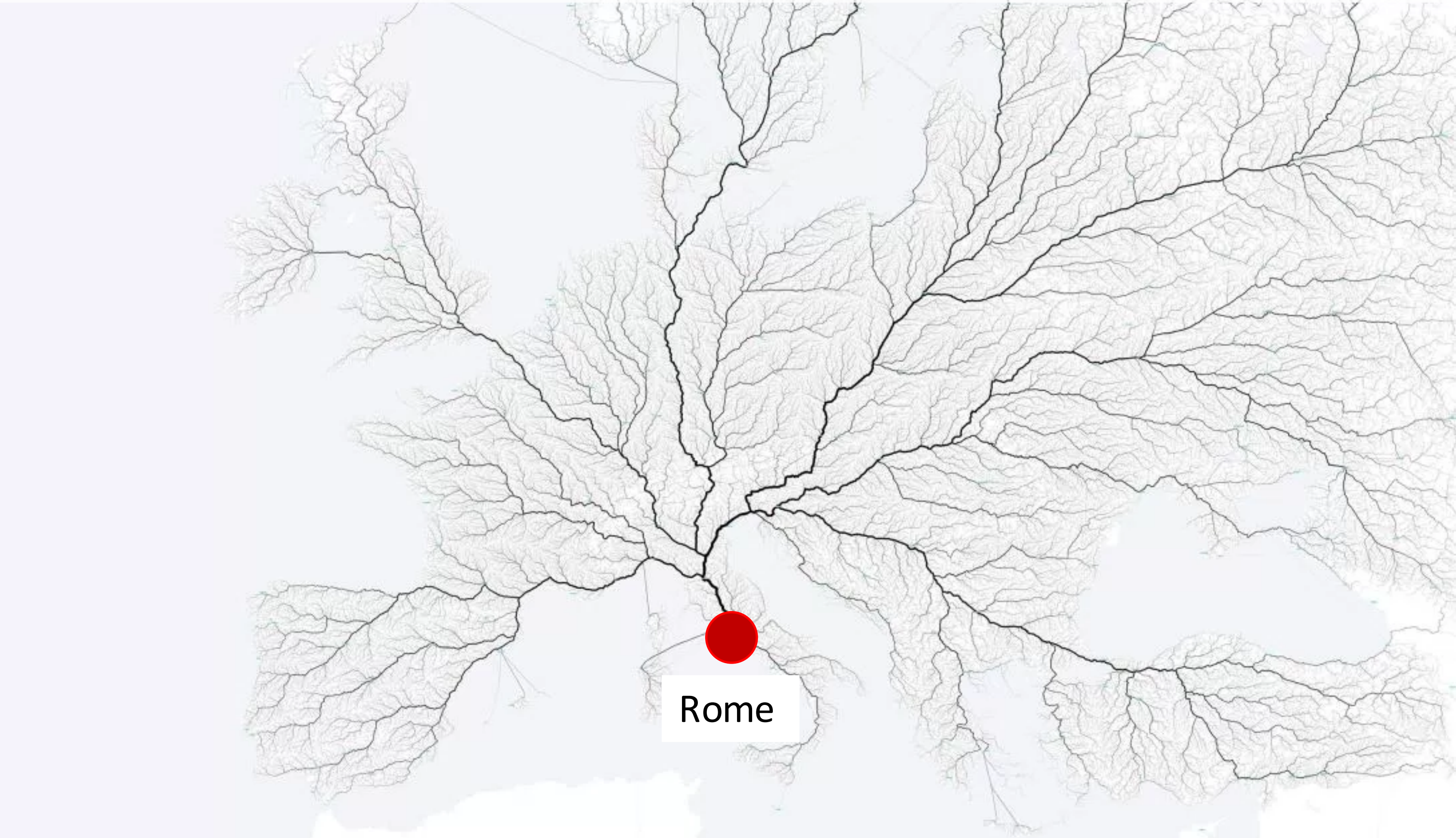
computation cost

Utility

frequent = interesting/useful?
no established metrics to evaluate pattern quality

Branching Pattern: an Inspiration

"All Roads to Rome: Visualizing Mobility at Scale", Reimann et. al. 2016

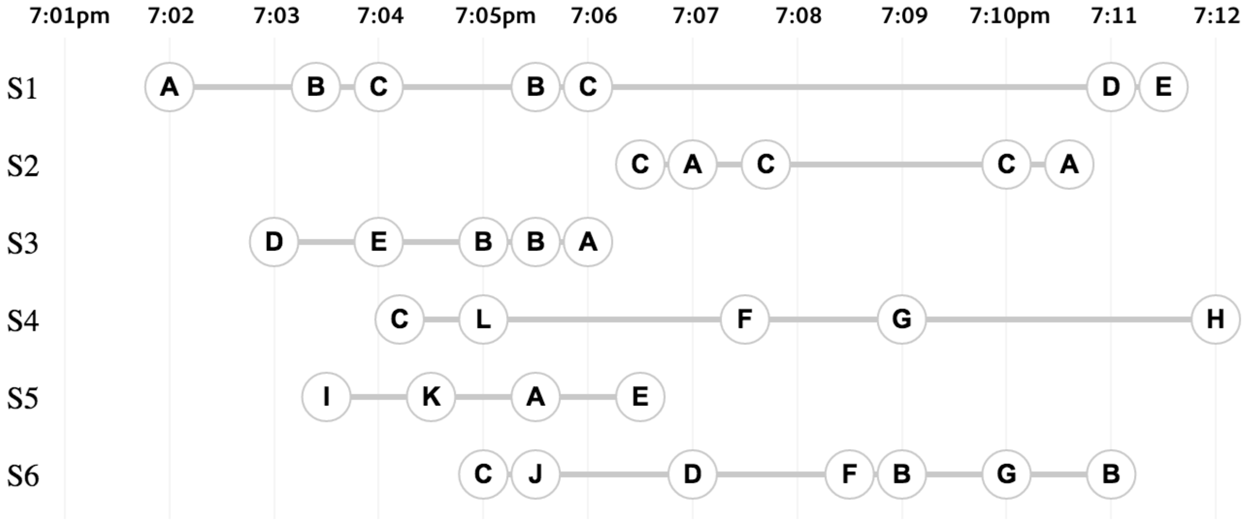


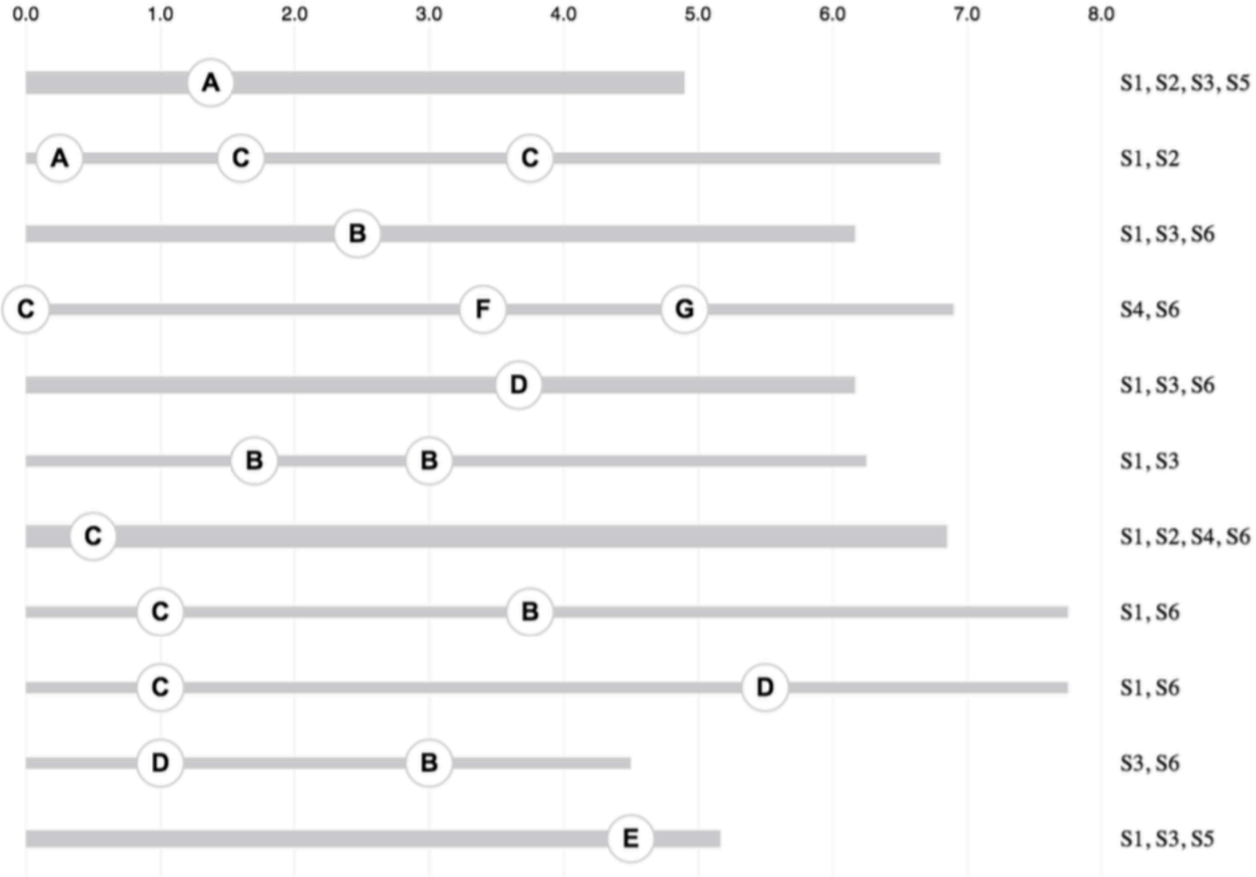
Branching Pattern: an Inspiration

Think of each sequence as a journey

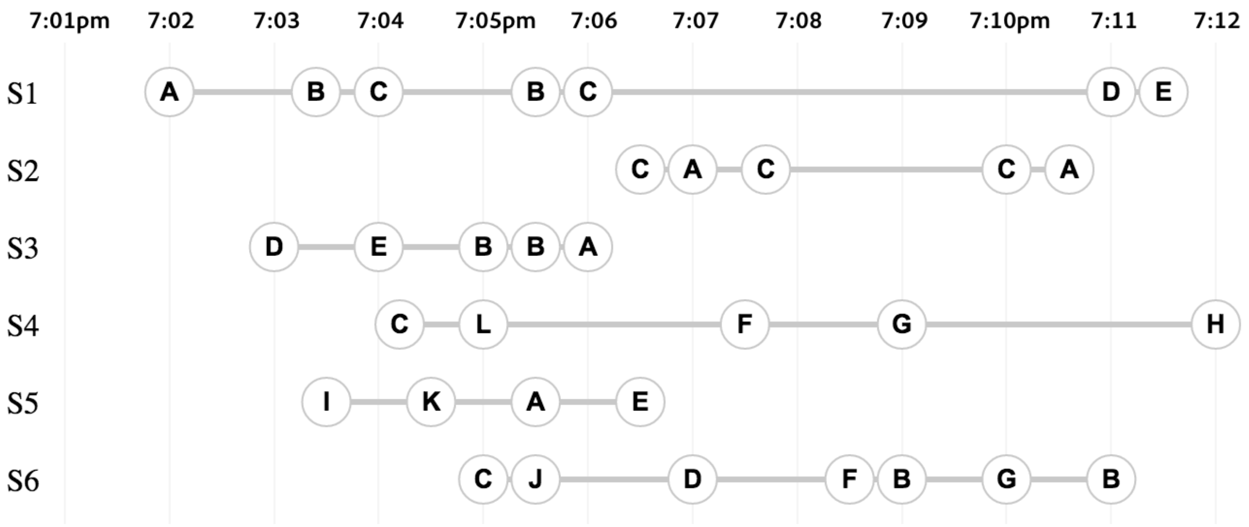
Despite differences in the exact paths and time taken,
the travelers may share a few common milestones in their journeys

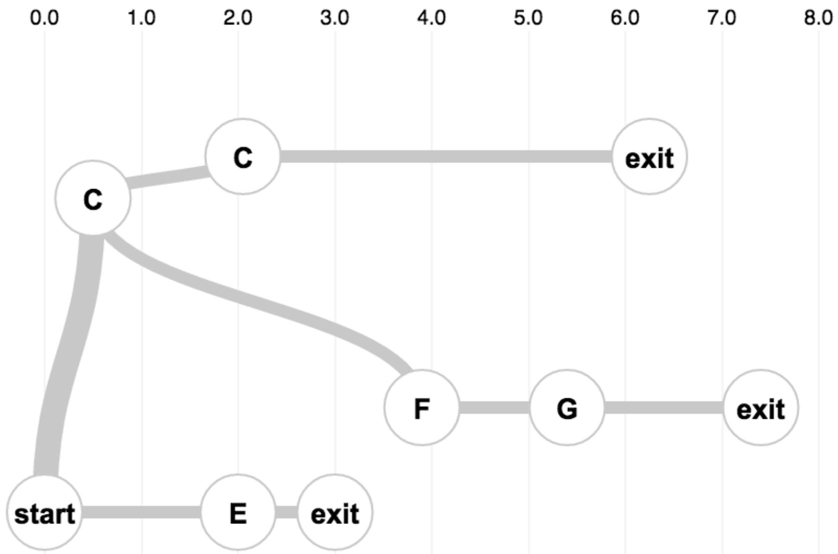
Sequential Pattern



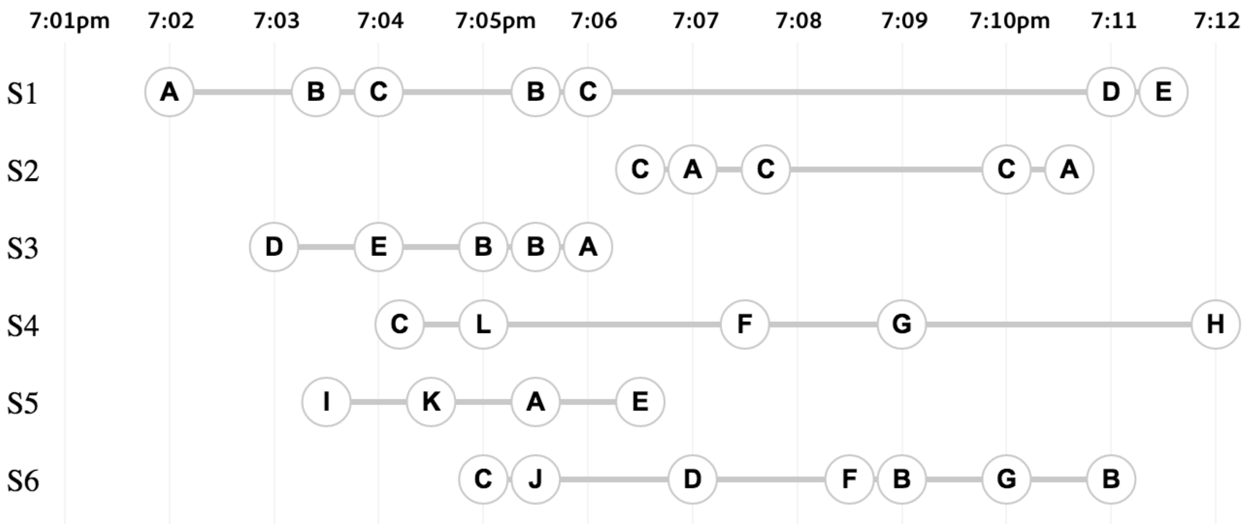


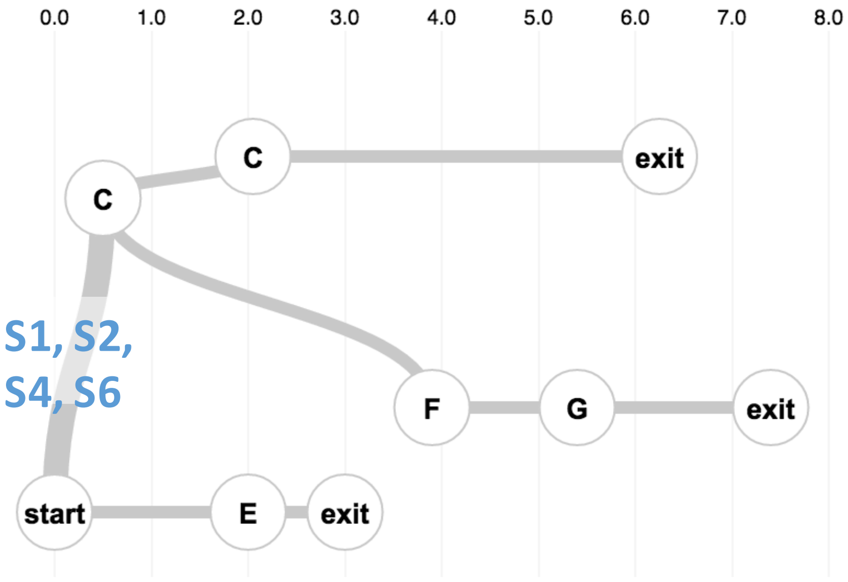
~~Sequential Pattern~~ Branching Pattern





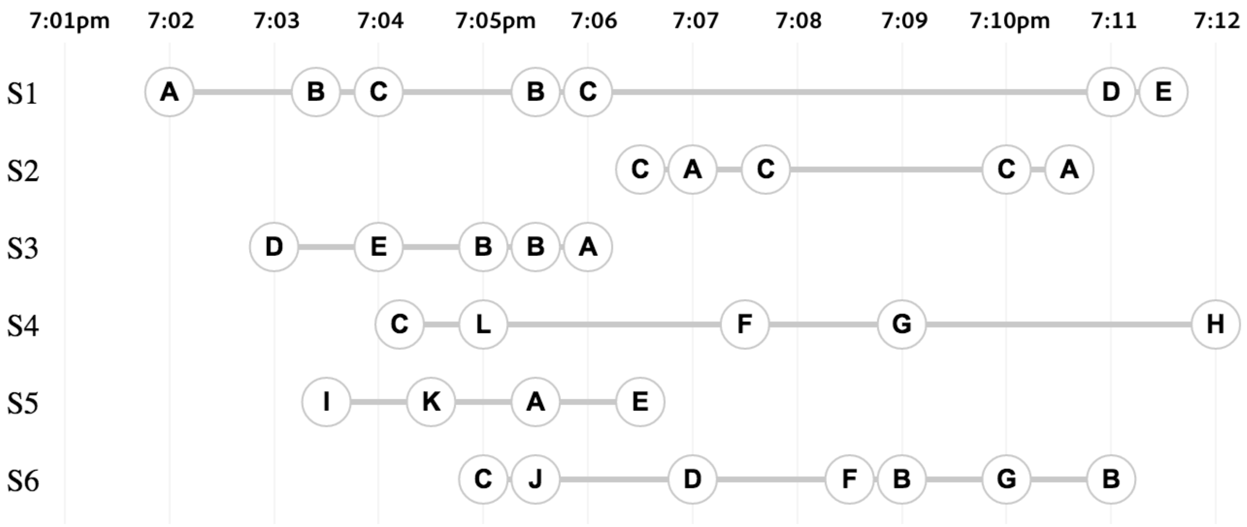
~~Sequential Pattern~~ Branching Pattern

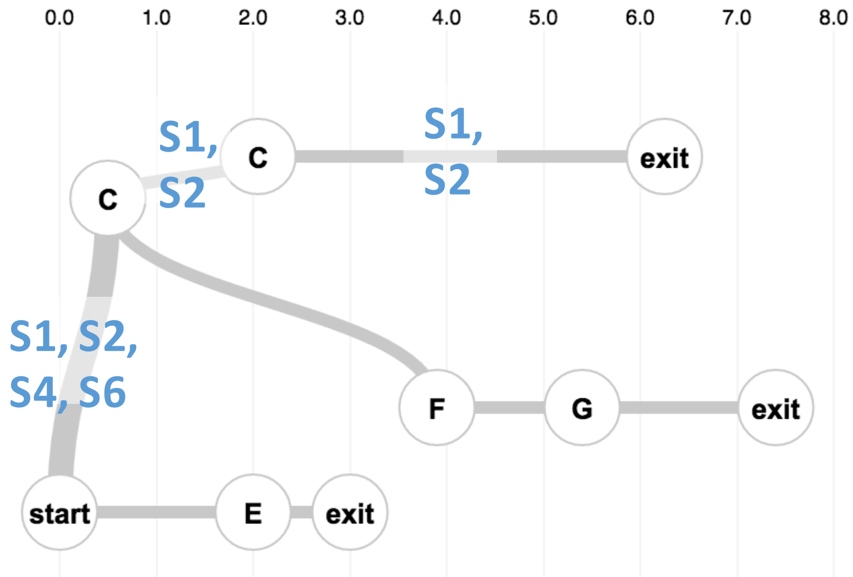




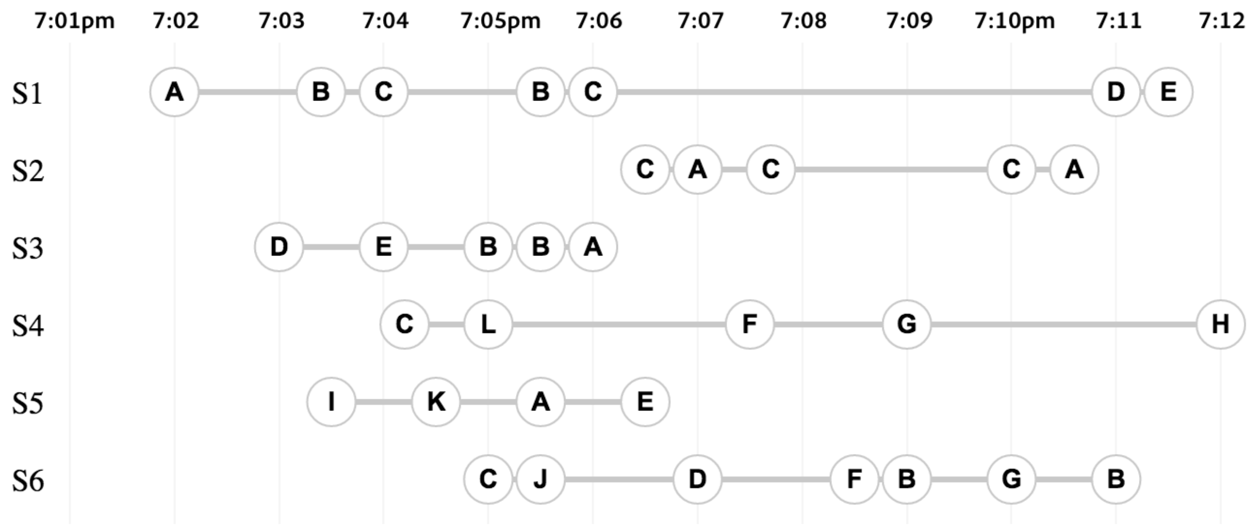
S1, S2,
S4, S6

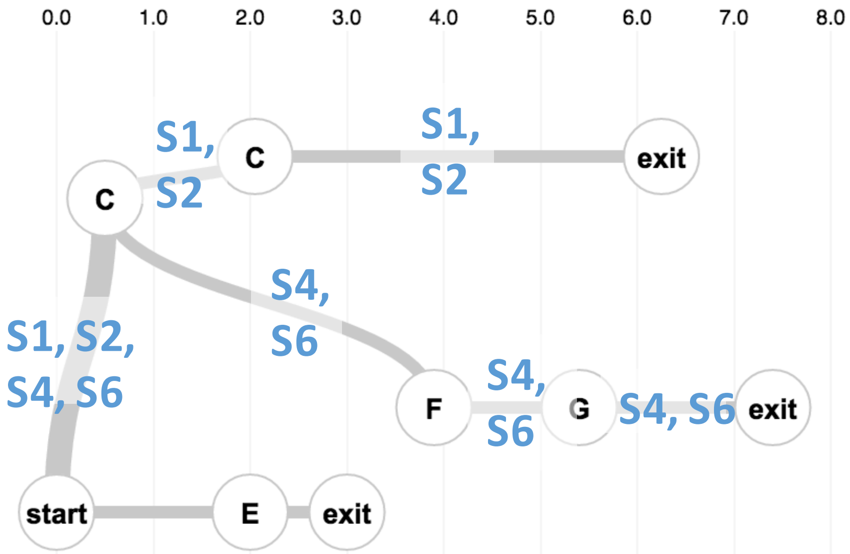
~~Sequential Pattern~~ Branching Pattern





~~Sequential Pattern~~ Branching Pattern





CoreFlow: Extract Branching Patterns

1. Rank events

2. Divide sequences

3. Trim sequences

Do this recursively until we run out of events or hit a predefined threshold

Input Sequences → Rank → Divide → Trim

ABCBCDE
CACCA
DEBBA
CLFGH
IKAE
CJDFBGB

EVT	# SEQ	AVG IDX
C	4	0.5
A	4	1.75
B	3	2.33
D	3	2.33
E	3	3.33

ABCBCDE
CACCA
CLFGH
CJDFBGB

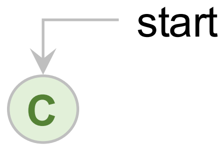
- - - - - - - - - - - - - - -

DEBBA
IKAE

~~ABC~~BCDE
~~C~~ACCA
~~C~~LFGH
~~C~~JDFBGB

- - - - - - - - - - - - - - -

DEBBA
IKAE



Input Sequences → Rank → Divide → Trim

~~ABC~~BCDE
~~C~~ACCA
~~C~~LFGH
~~C~~JDFBGB

- - - - - - - - - - - - - - -

DEBBA
IKAE

EVT	# SEQ	AVG IDX
C	2	1.0
D	2	1.5
F	2	1.5
B	2	2
G	2	3

~~ABC~~BCDE
~~C~~ACCA

- - - - - - - - - - - - - - -

~~C~~LFGH
~~C~~JDFBGB

- - - - - - - - - - - - - - -

DEBBA
IKAE

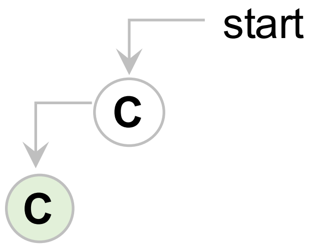
~~ABCBC~~DE
~~CAC~~CA

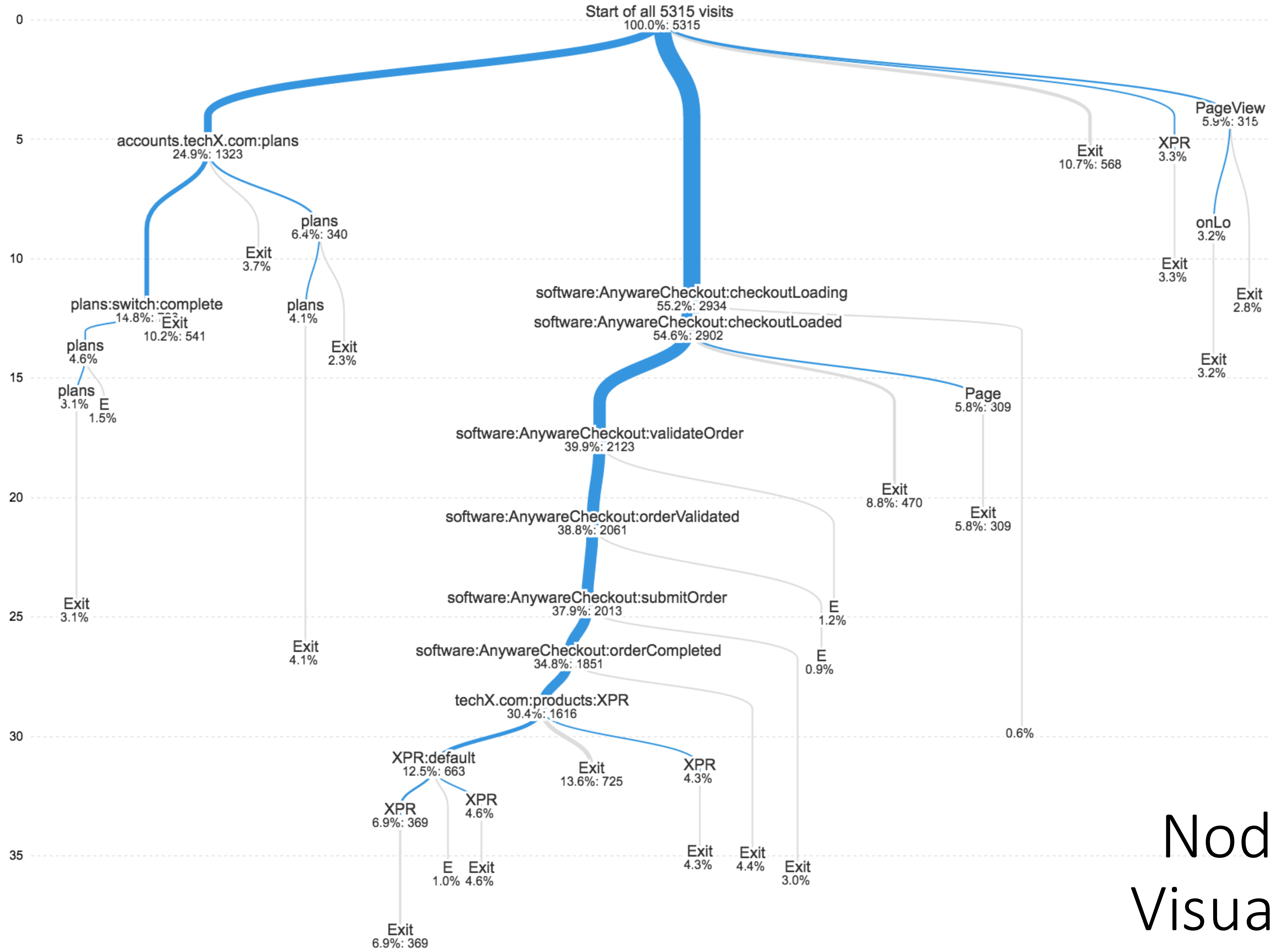
- - - - - - - - - - - - - - -

~~C~~LFGH
~~C~~JDFBGB

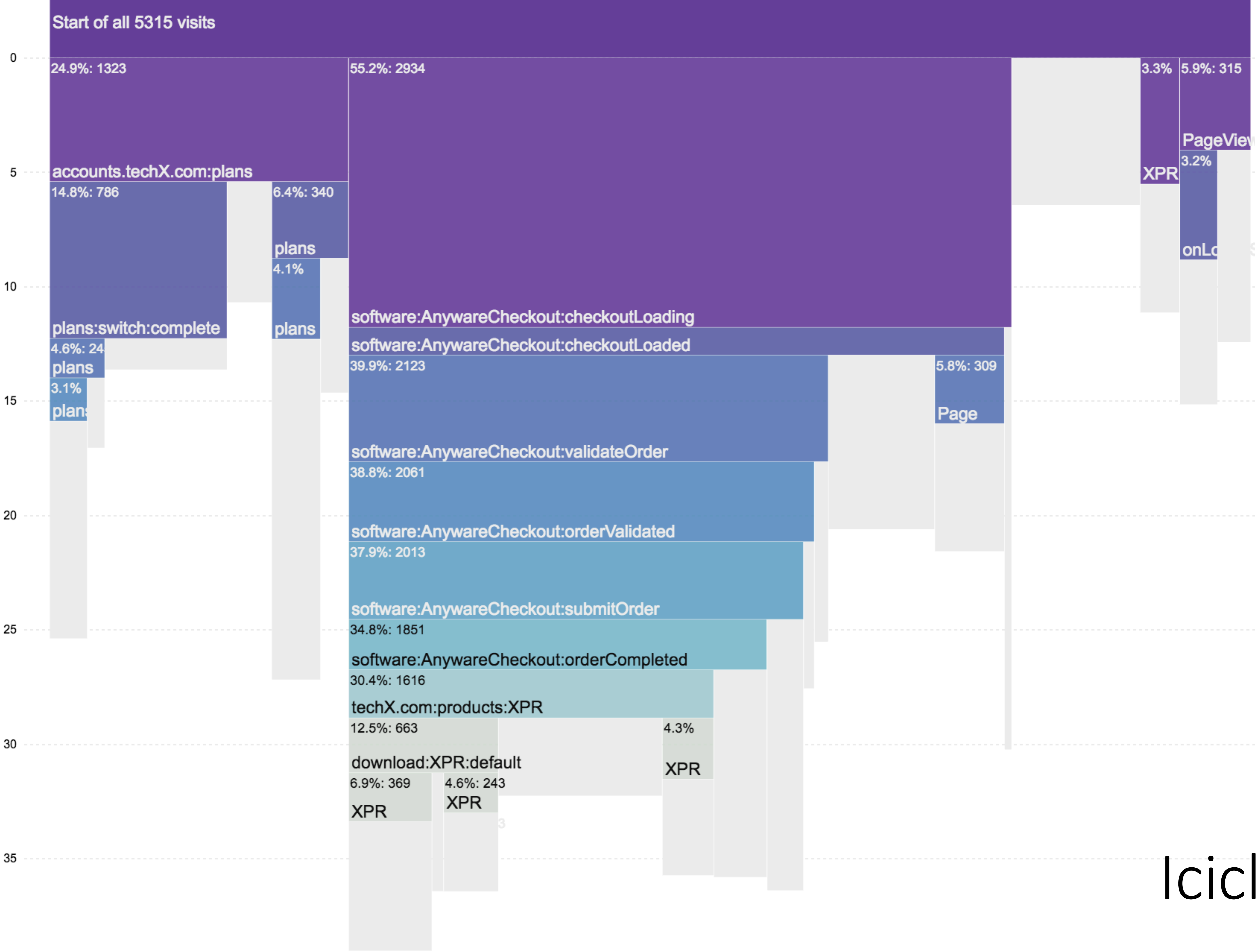
- - - - - - - - - - - - - - -

DEBBA
IKAE

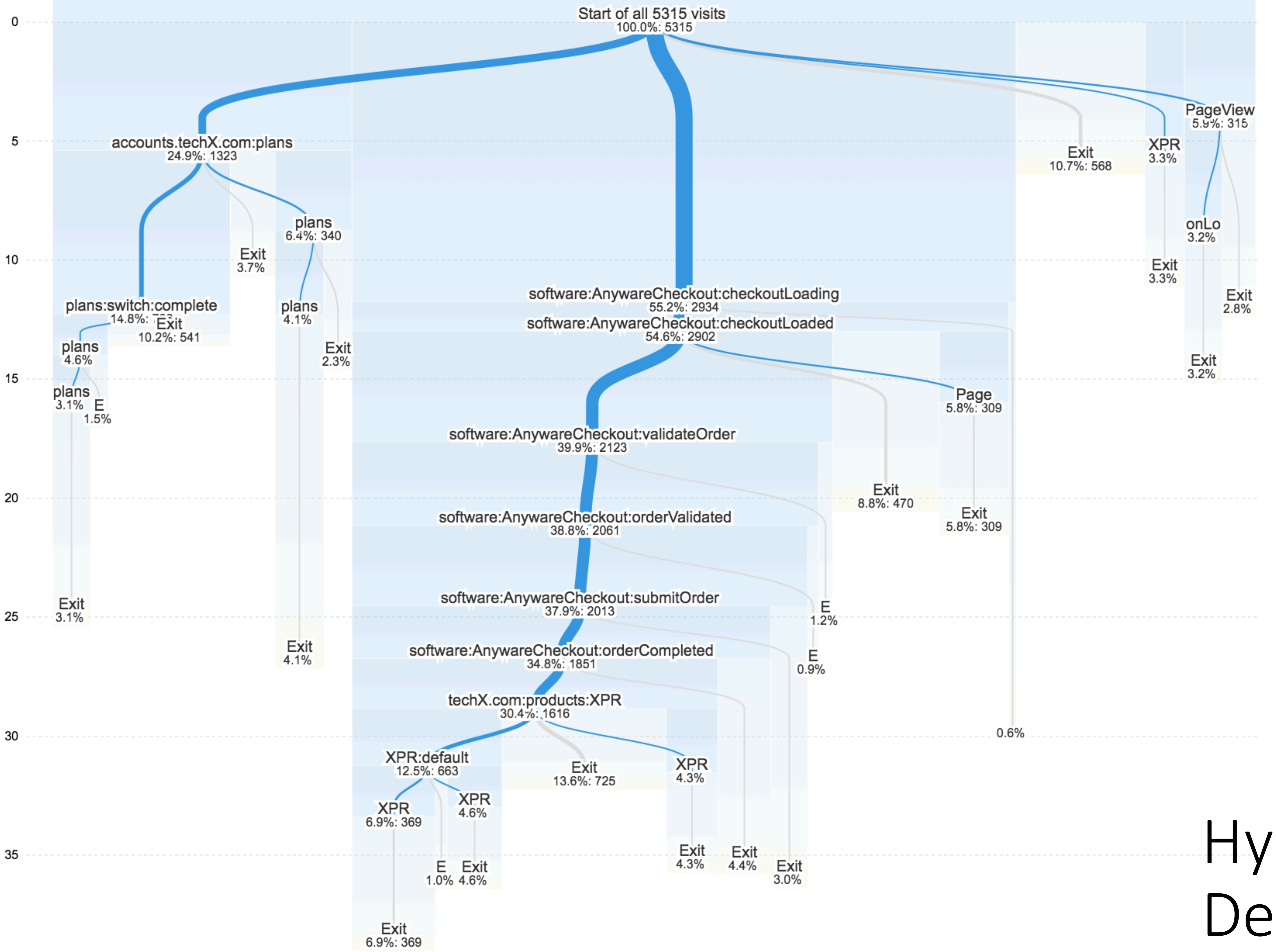




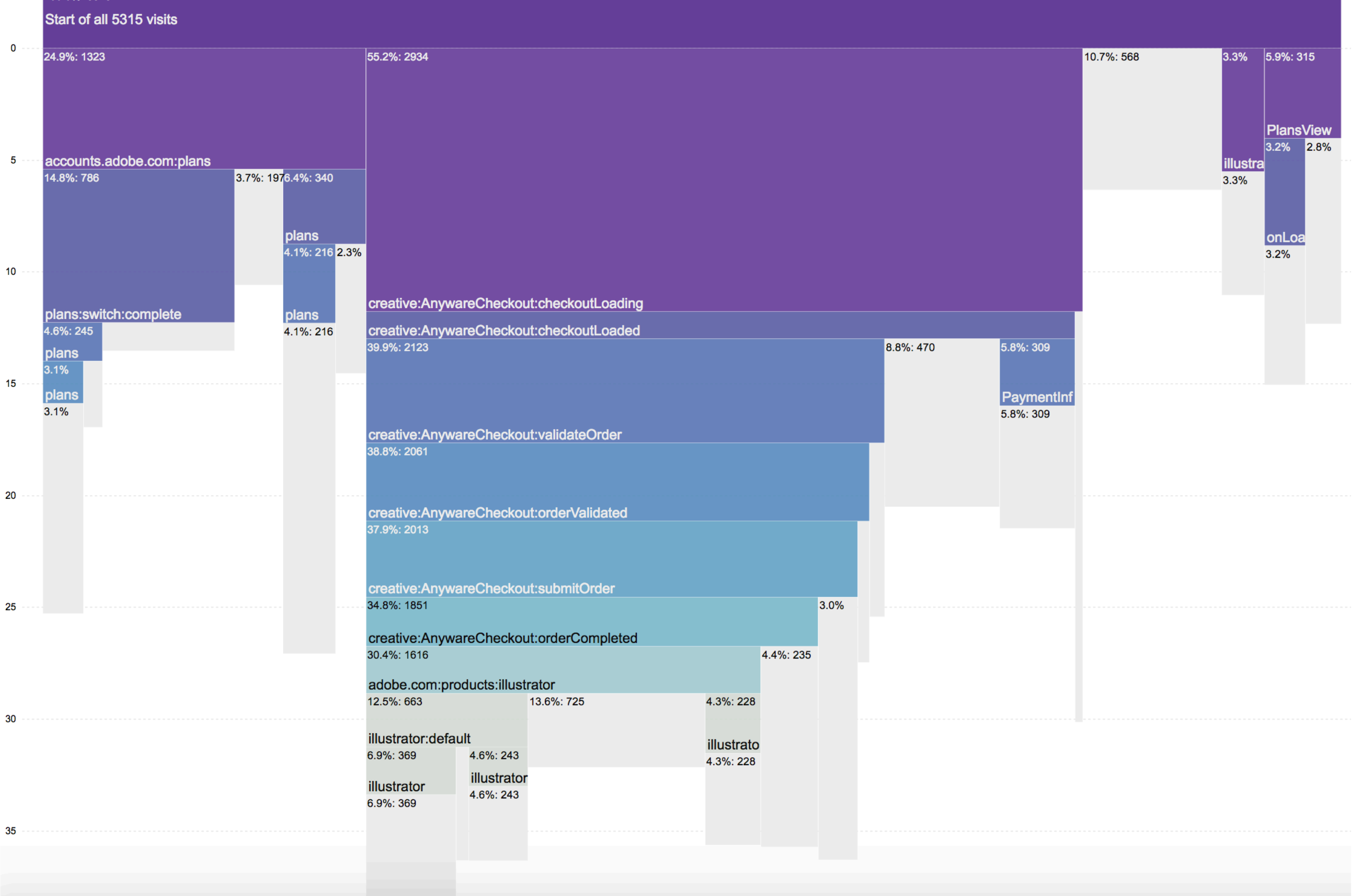
Node-Link Visualization

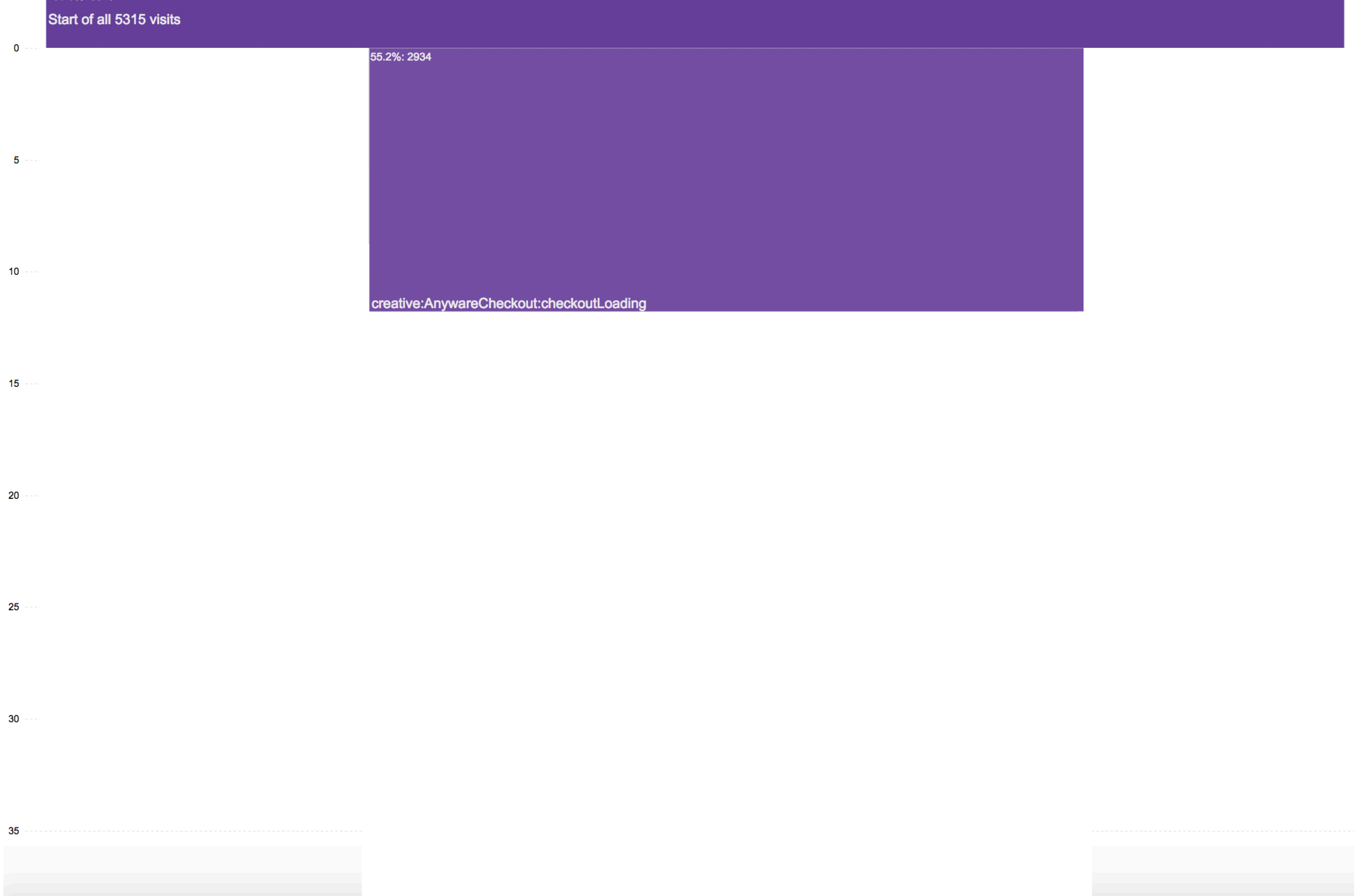

Start of all 5315 visits
0
5
10
15
20
25
30
35
24.9%: 1323
accounts.techX.com:plans
55.2%: 2934
software:AnywareCheckout:checkoutLoading
3.3%
XPR
5.9%: 315
PageView
3.2%
onLo
14.8%: 786
plans:switch:complete
6.4%: 340
plans
4.1%
plans
4.6%: 24
plans
3.1%
plans
software:AnywareCheckout:checkoutLoaded
39.9%: 2123
software:AnywareCheckout:validateOrder
5.8%: 309
Page
38.8%: 2061
software:AnywareCheckout:orderValidated
37.9%: 2013
software:AnywareCheckout:submitOrder
34.8%: 1851
software:AnywareCheckout:orderCompleted
30.4%: 1616
techX.com:products:XPR
12.5%: 663
download:XPR:default
4.3%
XPR
6.9%: 369
XPR
4.6%: 243
XPR

Icicle Plot

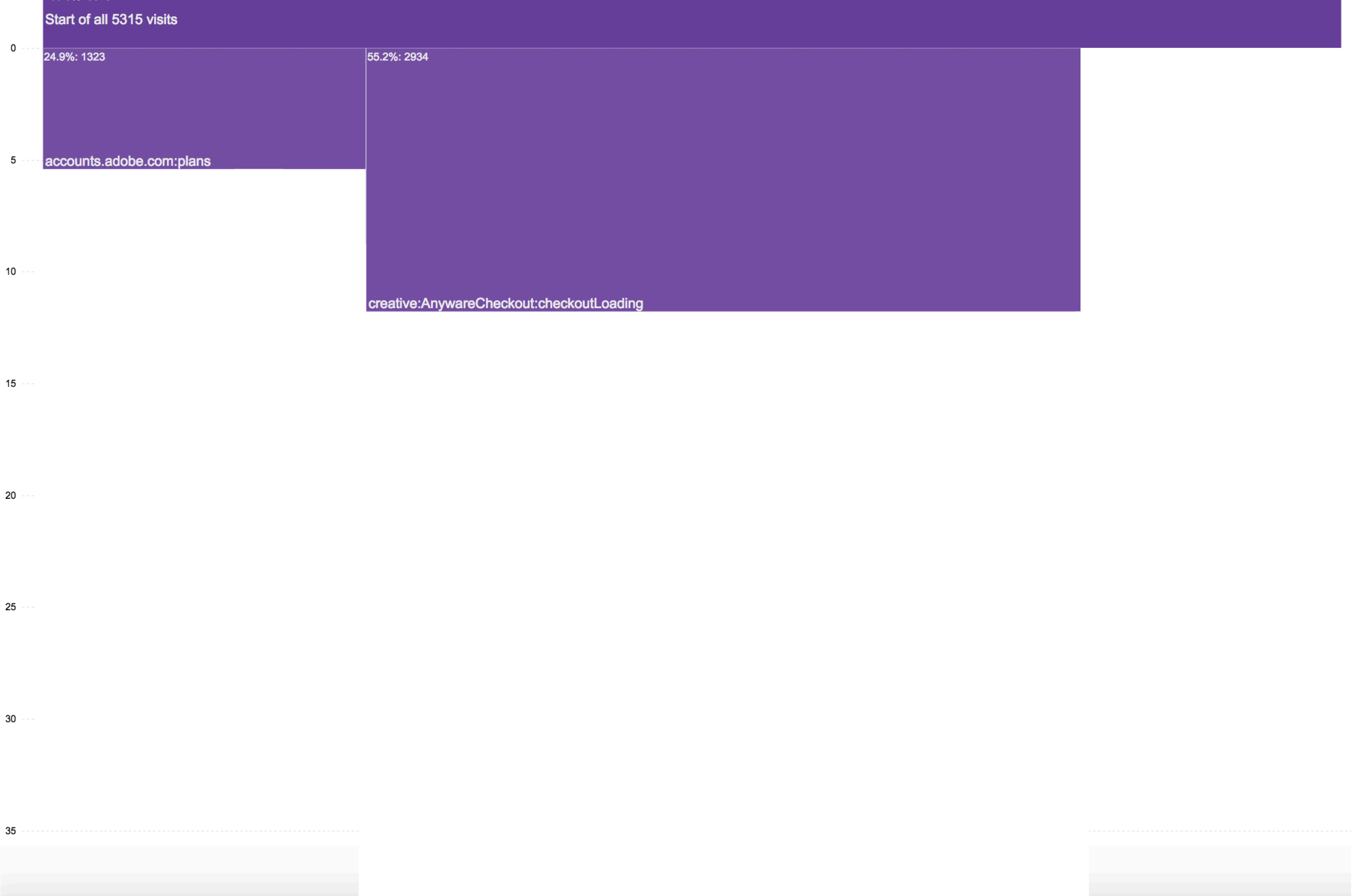


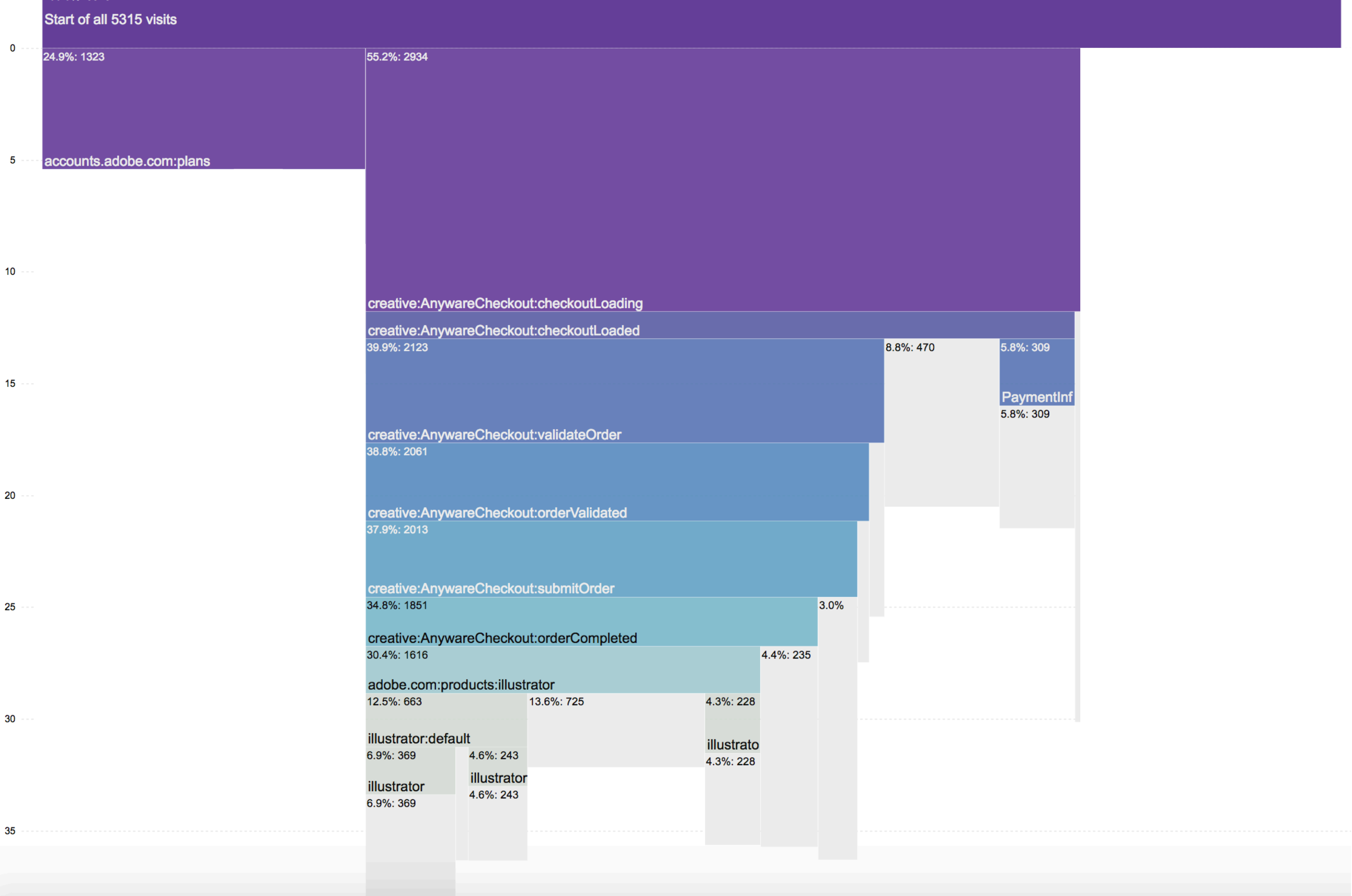
Hybrid Design


Start of all 5315 visits
0
5
10
15
20
25
30
35
24.9%: 1323
accounts.adobe.com:plans
14.8%: 786
plans:switch:complete
4.6%: 245
plans
3.1%
plans
3.1%
3.7%: 197
6.4%: 340
plans
4.1%: 216
plans
4.1%: 216
2.3%
55.2%: 2934
creative:AnywareCheckout:checkoutLoading
creative:AnywareCheckout:checkoutLoaded
39.9%: 2123
creative:AnywareCheckout:validateOrder
38.8%: 2061
creative:AnywareCheckout:orderValidated
37.9%: 2013
creative:AnywareCheckout:submitOrder
34.8%: 1851
creative:AnywareCheckout:orderCompleted
30.4%: 1616
adobe.com:products:illustrator
12.5%: 663
illustrator:default
6.9%: 369
illustrator
6.9%: 369
4.6%: 243
illustrator
4.6%: 243
13.6%: 725
4.3%: 228
illustrato
4.3%: 228
4.4%: 235
3.0%
8.8%: 470
5.8%: 309
PaymentInf
5.8%: 309
10.7%: 568
3.3%
illustra
3.3%
5.9%: 315
PlansView
3.2%
onLoa
3.2%
2.8%

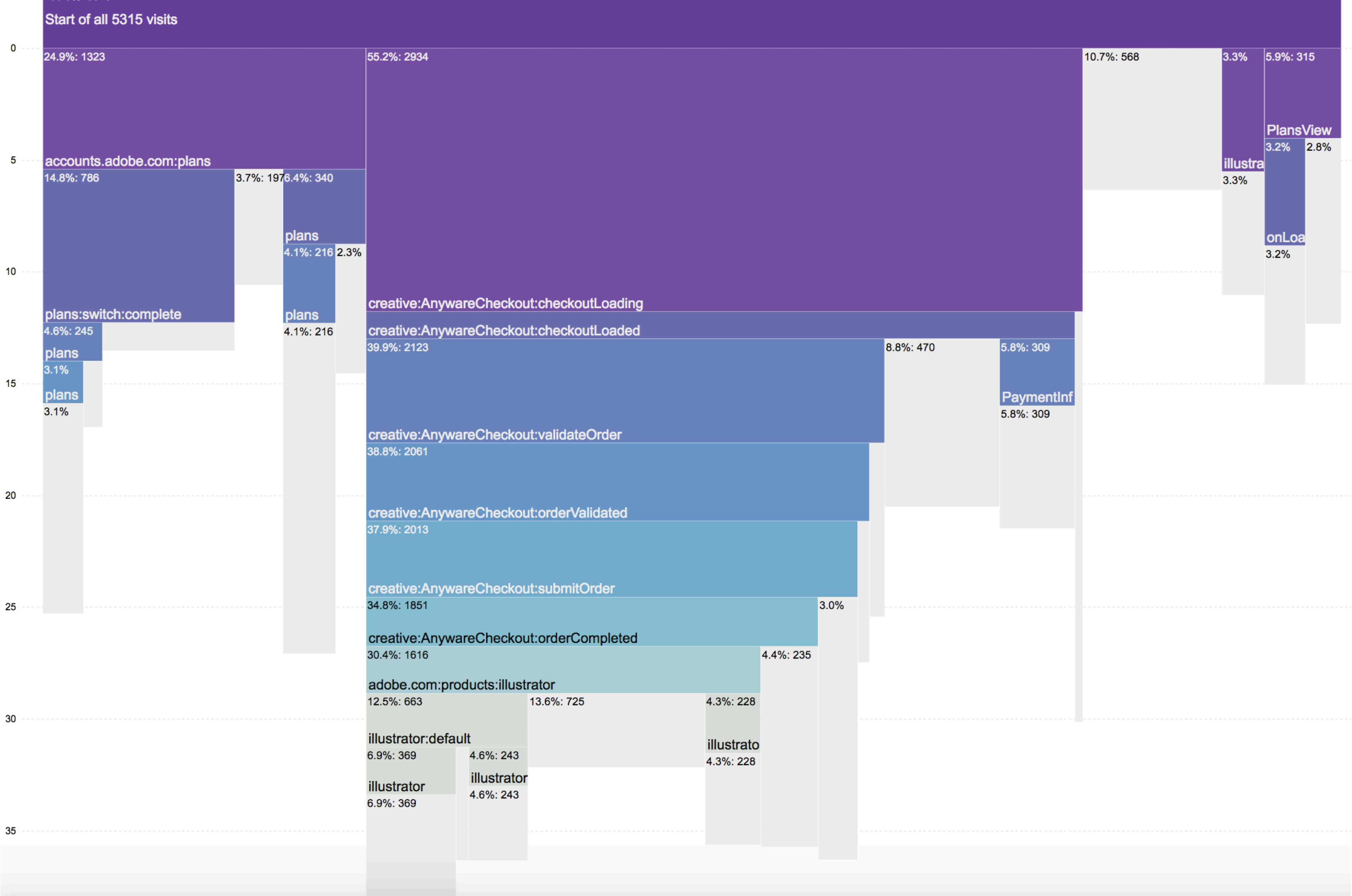


Start of all 5315 visits
55.2%: 2934
creative:AnywareCheckout:checkoutLoading
0
5
10
15
20
25
30
35

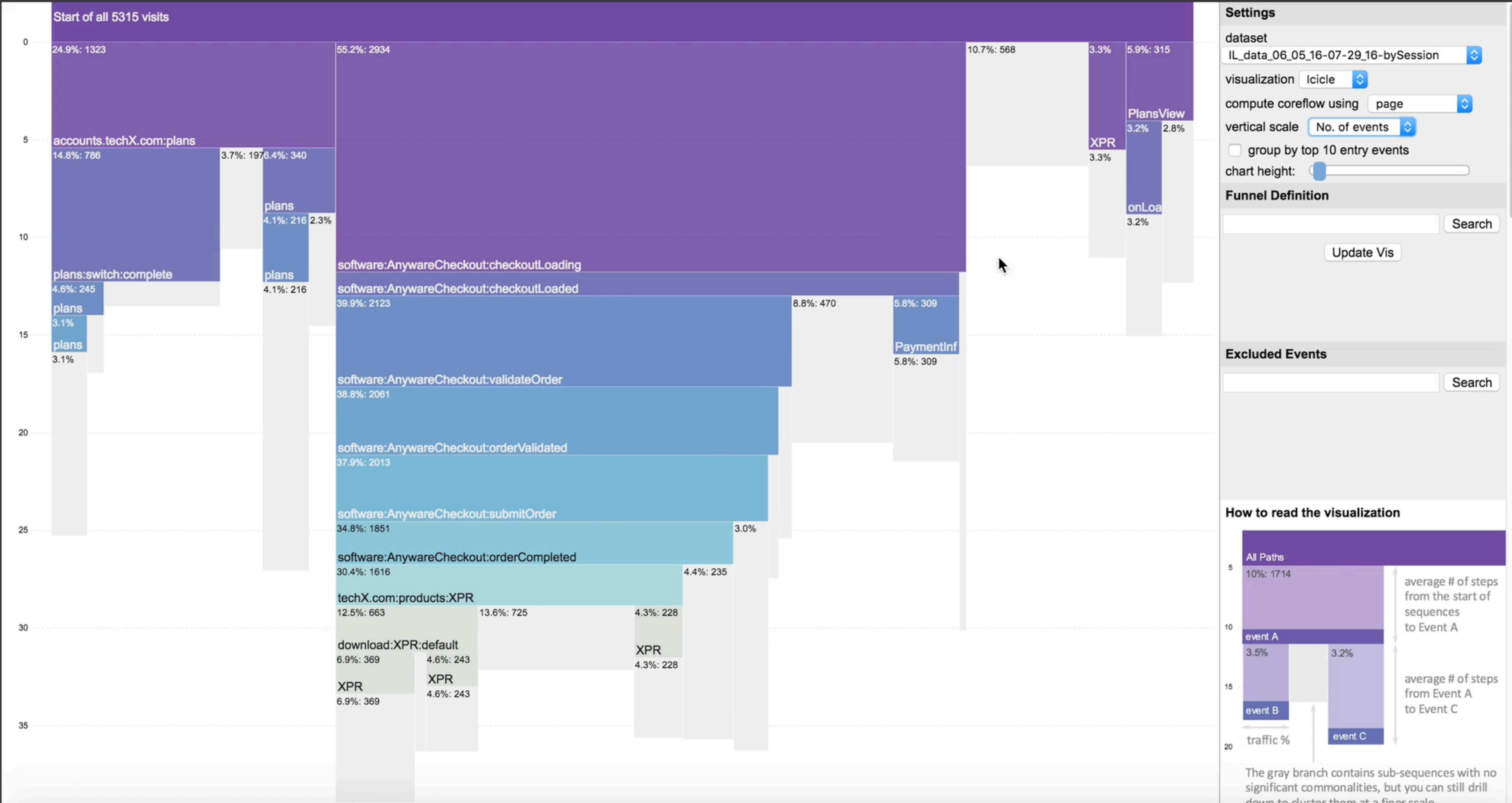

Start of all 5315 visits
24.9%: 1323
accounts.adobe.com:plans
55.2%: 2934
creative:AnywareCheckout:checkoutLoading
0
5
10
15
20
25
30
35



Start of all 5315 visits
0
5
10
15
20
25
30
35
24.9%: 1323
accounts.adobe.com:plans
55.2%: 2934
creative:AnywareCheckout:checkoutLoading
creative:AnywareCheckout:checkoutLoaded
39.9%: 2123
8.8%: 470
5.8%: 309
PaymentInf
5.8%: 309
creative:AnywareCheckout:validateOrder
38.8%: 2061
creative:AnywareCheckout:orderValidated
37.9%: 2013
creative:AnywareCheckout:submitOrder
34.8%: 1851
3.0%
creative:AnywareCheckout:orderCompleted
30.4%: 1616
4.4%: 235
adobe.com:products:illustrator
12.5%: 663
13.6%: 725
4.3%: 228
illustrator:default
illustrato
4.3%: 228
6.9%: 369
4.6%: 243
illustrator
illustrator
4.6%: 243
6.9%: 369



Start of all 5315 visits
0
5
10
15
20
25
30
35
24.9%: 1323
accounts.adobe.com:plans
14.8%: 786
plans:switch:complete
4.6%: 245
plans
3.1%
plans
3.1%
3.7%: 197
6.4%: 340
plans
4.1%: 216
plans
4.1%: 216
2.3%
55.2%: 2934
creative:AnywareCheckout:checkoutLoading
creative:AnywareCheckout:checkoutLoaded
39.9%: 2123
creative:AnywareCheckout:validateOrder
38.8%: 2061
creative:AnywareCheckout:orderValidated
37.9%: 2013
creative:AnywareCheckout:submitOrder
34.8%: 1851
creative:AnywareCheckout:orderCompleted
30.4%: 1616
adobe.com:products:illustrator
12.5%: 663
illustrator:default
6.9%: 369
illustrator
6.9%: 369
4.6%: 243
illustrator
4.6%: 243
13.6%: 725
4.3%: 228
illustrato
4.3%: 228
4.4%: 235
3.0%
8.8%: 470
5.8%: 309
PaymentInf
5.8%: 309
10.7%: 568
3.3%
illustra
3.3%
5.9%: 315
PlansView
3.2%
onLoa
3.2%
2.8%



Settings

dataset IL_data_06_05_16-07-29_16-bySession

visualization Icicle

compute coreflow using page

vertical scale No. of events

□ group by top 10 entry events

chart height:

Funnel Definition

Search

Update Vis

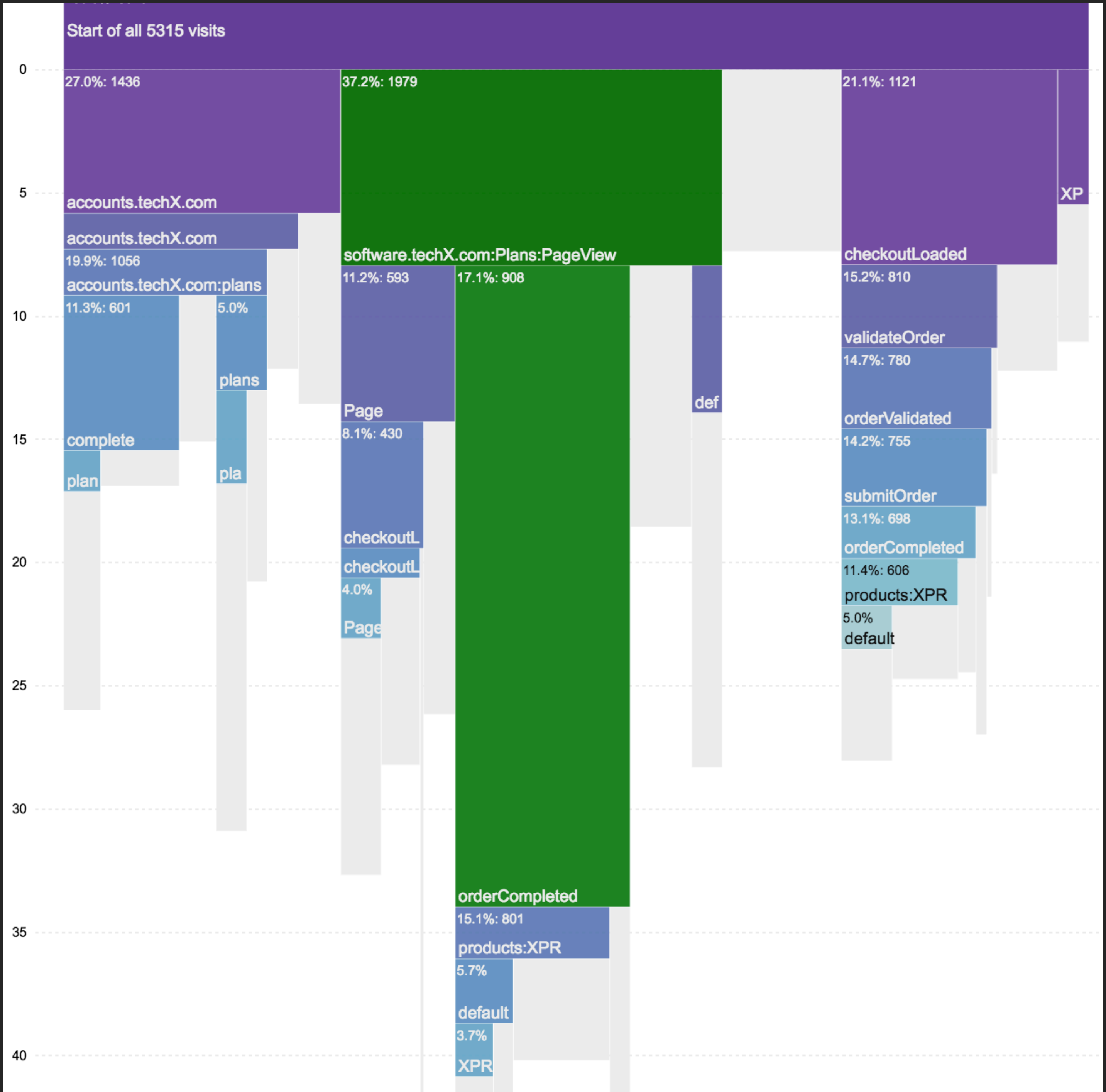
Excluded Events

Search

How to read the visualization

All Paths
5
10
15
20
10%: 1714
event A
3.5%
3.2%
event B
event C
average # of steps from the start of sequences to Event A
average # of steps from Event A to Event C
traffic %

The gray branch contains sub-sequences with no significant commonalities, but you can still drill down to cluster them at a finer scale


Start of all 5315 visits
0
5
10
15
20
25
30
35
40
27.0%: 1436
accounts.techX.com
accounts.techX.com
19.9%: 1056
accounts.techX.com:plans
11.3%: 601
complete
plan
5.0%
plans
pla
37.2%: 1979
software.techX.com:Plans:PageView
11.2%: 593
Page
8.1%: 430
checkoutL
checkoutL
4.0%
Page
17.1%: 908
orderCompleted
15.1%: 801
products:XPR
5.7%
default
3.7%
XPR
def
21.1%: 1121
checkoutLoaded
15.2%: 810
validateOrder
14.7%: 780
orderValidated
14.2%: 755
submitOrder
13.1%: 698
orderCompleted
11.4%: 606
products:XPR
5.0%
default
XP

How does CoreFlow perform?

Interpretability

relationship between patterns
number of patterns to review

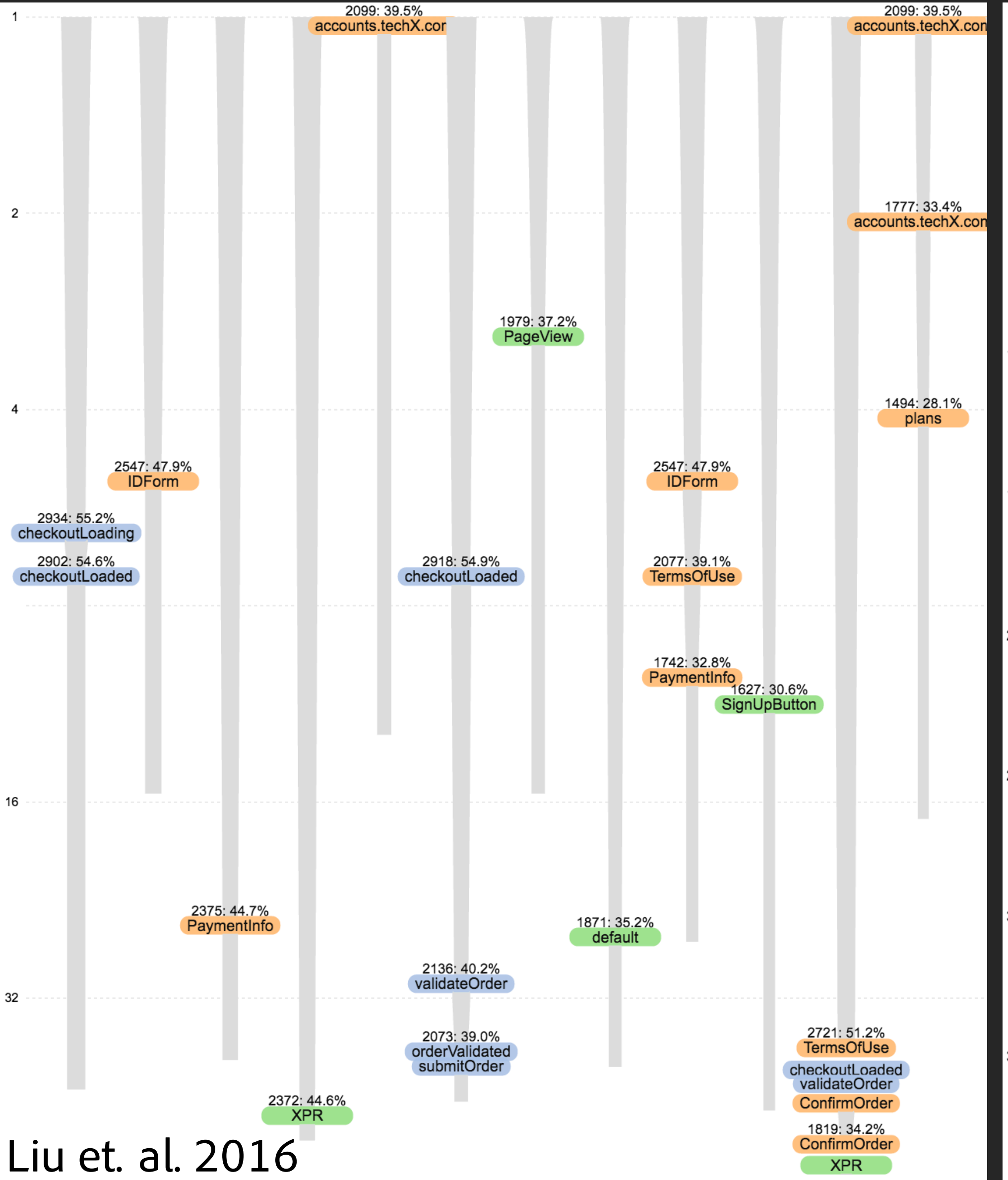
Scalability

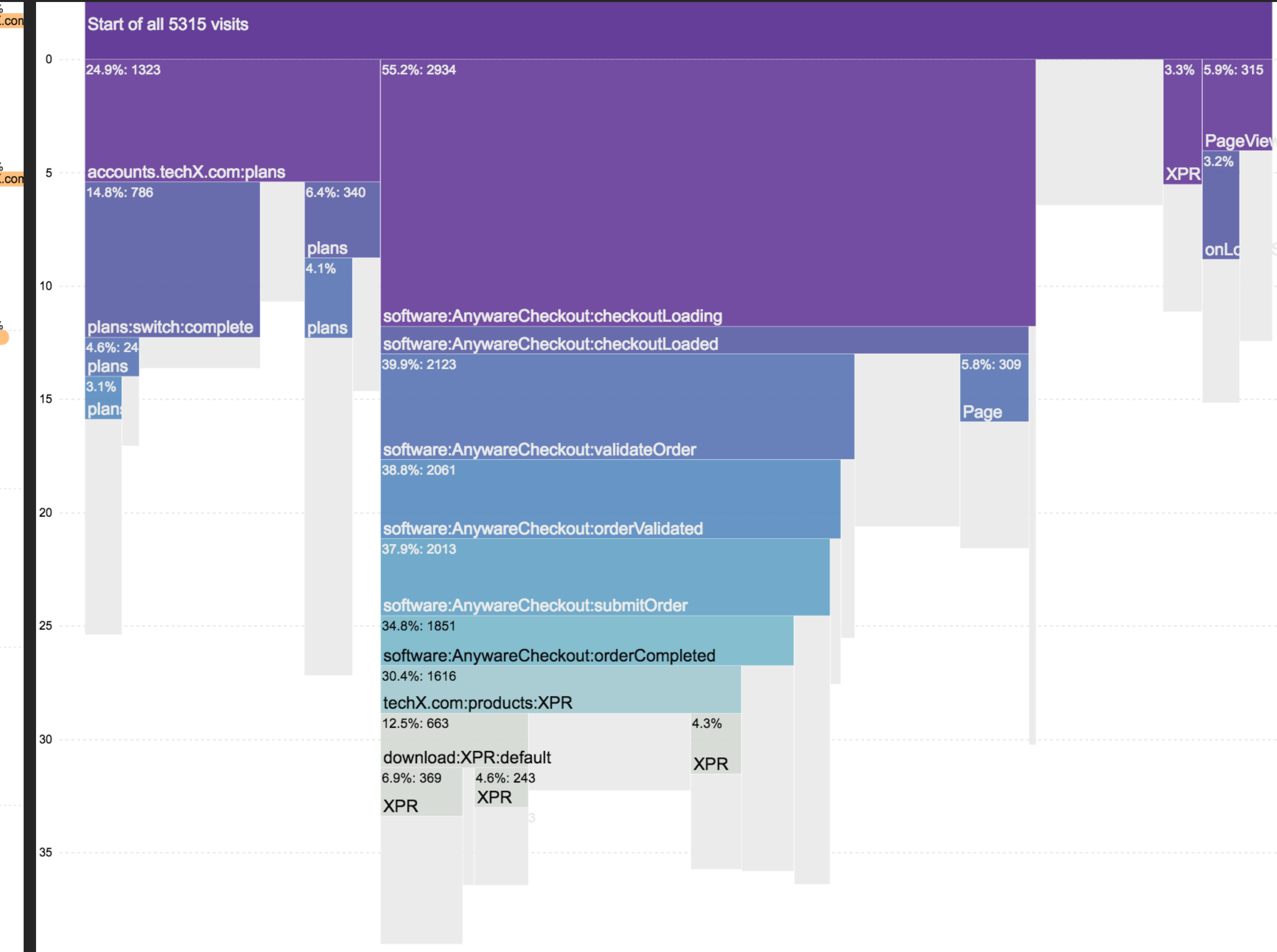
computation cost

Utility

frequent = interesting/useful?
no established metrics to evaluate pattern quality

Interpretability





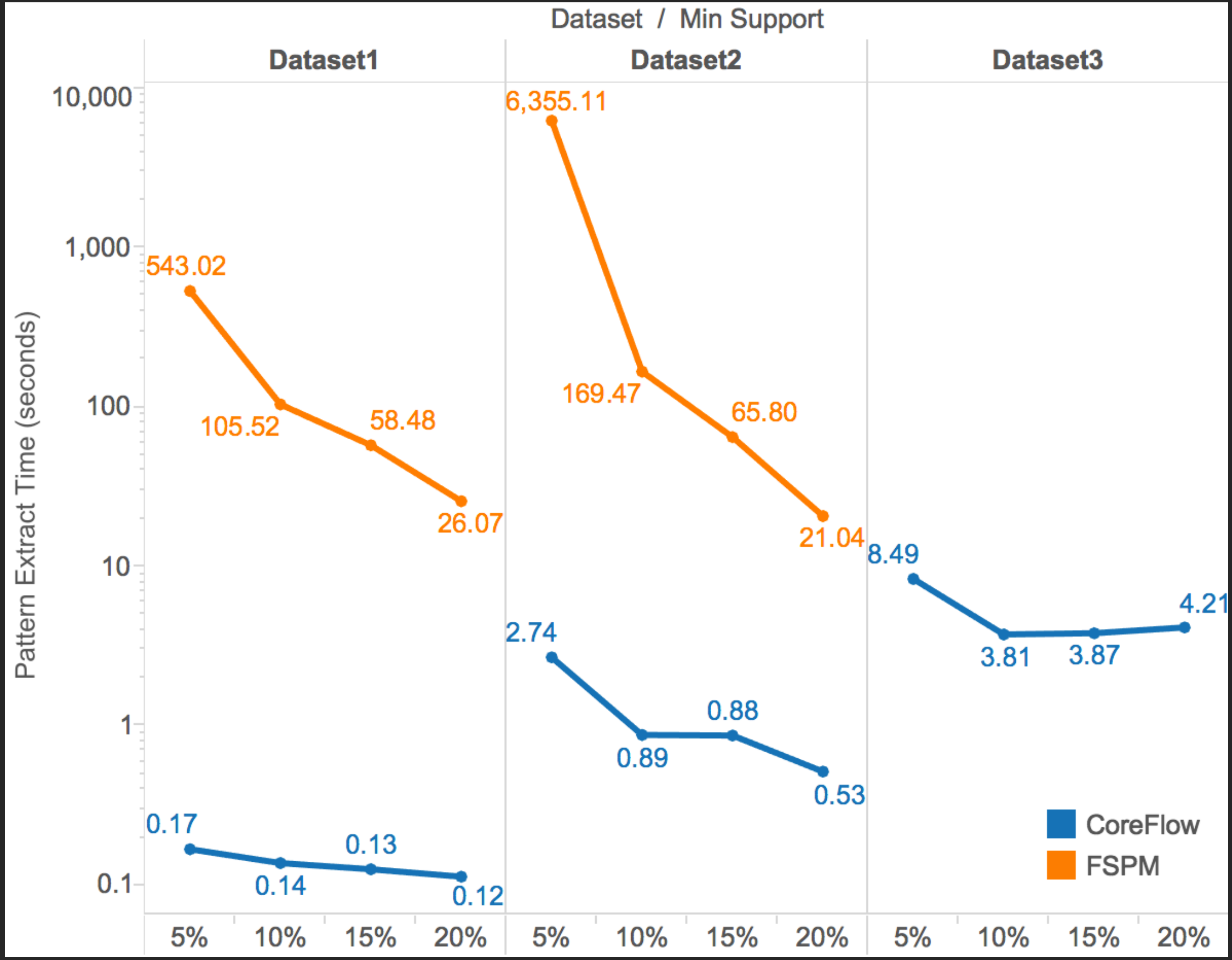
Liu et. al. 2016

Scalability

	Dataset 1	Dataset 2	Dataset 3
sequences	5K	60K	301K
events	121K	1.49M	5.5M
unique events	4.2K	36	107.7K
avg seq length	22.7	24.6	18.3
max seq length	186	3.2K	2.5K

Scalability

	Dataset 1	Dataset 2	Dataset 3
sequences	5K	60K	301K
events	121K	1.49M	5.5M
unique events	4.2K	36	107.7K
avg seq length	22.7	24.6	18.3
max seq length	186	3.2K	2.5K



Utility: Does frequency imply usefulness?

Methodology: Case study approach

use the analysts' domain knowledge as a baseline to evaluate the patterns

Three case studies in different domains

visitor paths in web clickstreams
workflows in application logs
touch points in marketing event paths

Varied level of success

Visitor Paths in Web Clickstreams

successfully identify meaningful milestone events

“this is perfect”, “potential to be very powerful”

Varied level of success

Visitor Paths in Web Clickstreams

successfully identify meaningful milestone events
“this is perfect”, “potential to be very powerful”

Workflows in Application Logs

a milestone is not an event, but a task
need to segment logs into meaningful tasks

Varied level of success

Visitor Paths in Web Clickstreams

successfully identify meaningful milestone events

“this is perfect”, “potential to be very powerful”

Workflows in Application Logs

a milestone is not an event, but a task

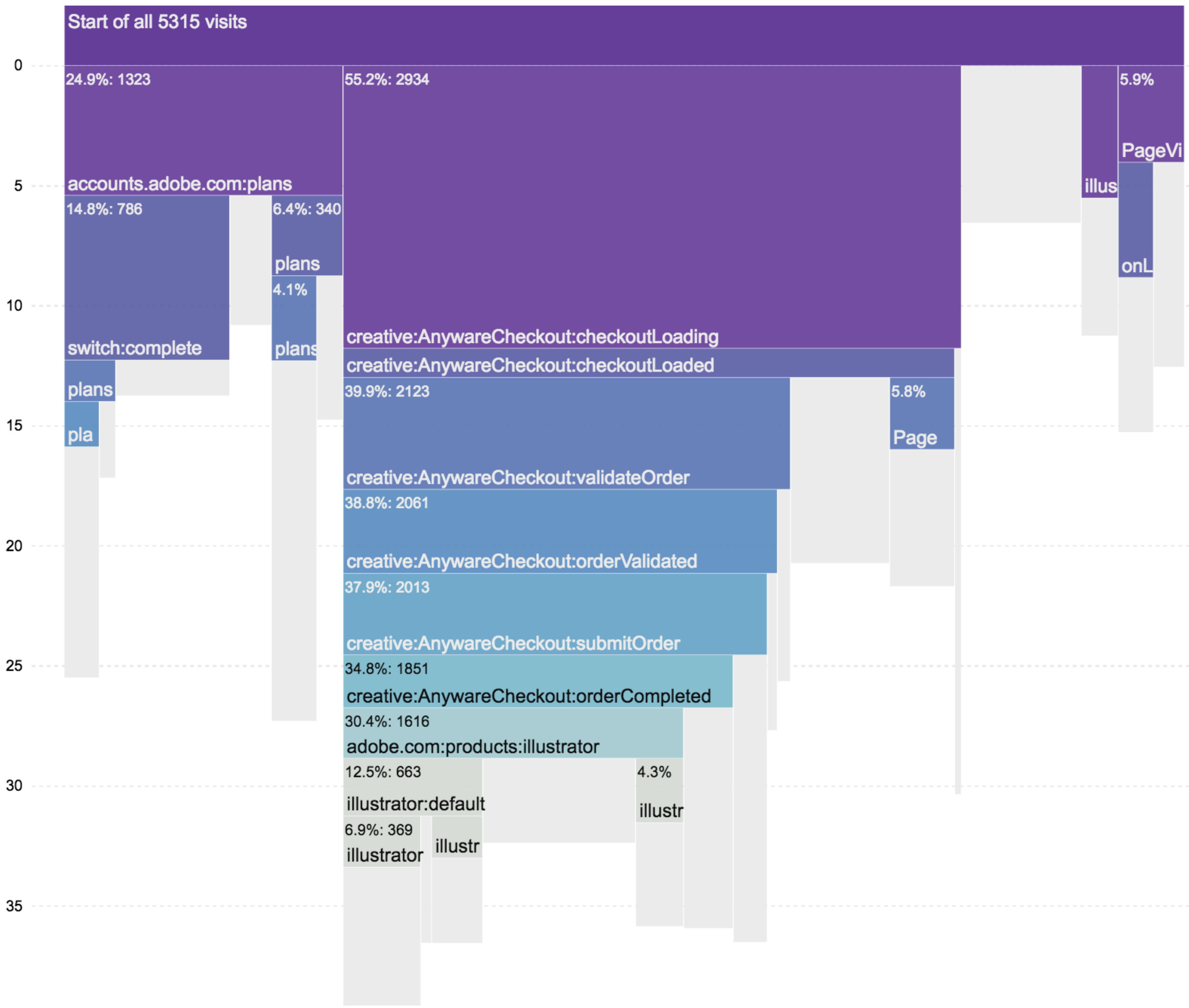
need to segment logs into meaningful tasks

```
Crop              Crop
New document      New document
New               New
Paste             Paste
Crop              Crop
New               New
Open              Open
New document      New document
Paste             Paste
Crop              Crop
```

Future Work

Evaluation metrics for pattern quality

Deeper understanding of the effects of ranking function on pattern quality



Supplemental Materials:

Video, Algorithms & Case Studies

http://www.zcliu.org/coreflow

Adobe