PNNL-SA-148524

VAST 2020 Preview

The VAST challenge committee is pleased to provide a preview of the 2020 VAST challenge. Like Mini-Challenge 3 in 2018 (http://www.vacommunity.org/VAST+Challenge+2018+MC3), one of the mini-challenges in 2020 will revolve around transactional records that can be represented, analyzed, and visualized as a graph. Next year's challenge will increase the level of difficulty by increasing the number of data channels and the number of records to be analyzed. As a result, we are aiming to give participants a head start by first becoming familiar with the 6 distinct data channels that will be release later this year. As a sample, we are providing:

- 100k email metadata records covering about 9 hours
- 100k call metadata records, covering about 5.5 hours
- 1000 procurement records covering about 5 hours
- 1000 co-authorship records covering a period from 85 to 45 years before the other channels start
- Demographics records for ~500 people
- 1000 travel records covering 2 days

Each dataset has a list of transactions that can be thought of as 'edges' on the graph.

Source	eType	Target	Time	Weight
42512	0	46456	86400	1
46456	0	47543	86400	1
47543	0	42512	86400	1
35957	0	17735	86400	1
76868	0	35957	86400	1
28161	0	39587	86400	1
39587	0	51558	86400	1
51558	0	28161	86400	1
20088	0	38632	86400	1
19897	0	62086	86400	1
62086	0	58476	86400	1
58476	0	19897	86400	1

Most data channels have 5 data columns where Source is the originator of the transaction and destination is the recipient. Time is in seconds from an arbitrary starting point in the near future (similar to Unix timestamps, except that time=0 is around the year 2020, and not 1970). All channels have a column for Weight, but it not used in the call or email channel. eType indicates the type of transaction.

Edge types are as follows:

Email
Phone
Sell (procurement)
Buy (procurement)
Author-of

Financial (income or expenditure, depending on direction)
Travels-to

Node types are as follows:

Person (used in all channels)
Product category (from the procurement graph)
Document (from the co-authorship graph)
Financial category (from financial demographics)
Country (from the travel graph)

Call and travel records have 6 additional columns for the location of the source and target.

Source	eType	Target	Time	Weight	SourceLocation	TargetLocation	SourceLatitude	SourceLongitude	TargetLatitude	TargetLongitude
175365	1	122860	86400	1	0	2	38.944	-37.9047	-20.4103	91.3391
122860	1	103377	86400	1	2	1	-20.4103	91.3391	-29.3767	-11.7244
103377	1	175365	86400	1	1	0	-29.3767	-11.7244	38.944	-37.9047
127598	1	160199	86400	1	5	0	20.3977	154.33	41.8345	-42.689
137341	1	176609	86400	1	3	4	-25.0473	-111.633	2.57221	-166.804
176609	1	174320	86400	1	4	3	2.57221	-166.804	-24.1415	-111.051
117303	1	155825	86400	1	4	0	5.50556	-164.253	27.7967	-34.4573

The SourceLocation and TargetLocation are integer values between 0 and 5, representing fictitious countries. The Latitude and Longitude columns are locations within the country associated with the person. There may be a significant amount of noise associated with this location.

If you have any questions, please contact the VAST Challenge committee at VASTChallenges@gmail.com or via GitHub (https://github.com/vast-challenge/2020-sample-data). We may ask to post your question anonymously to GitHub so others can see the response.

We are looking forward to enriching this big graph dataset and creating a unique challenge for visual analysis in 2020. We'll aim to keep the format of the data as close to the format of these samples as possible so you won't have to start from scratch on your tools and workflows for handling the data.