Uberlink Twitter Analytics Service (TAS) – Demonstration Dataset

Robert Ackland* Jamsheed Shorish†

April 6, 2017
Document version: 1.0
Data file format version: 1.0

This document describes the Uberlink Twitter Analytics Service (TAS) demonstration dataset, and associated Perl and R scripts for working with the data. The dataset and code are designed to assist Uberlink TAS users to understand the data structure and quickly start working with the data.

The demonstration dataset is constructed from Twitter "activities" (tweets/retweets) that are either authored by a sample of Australian advocacy organizations or else reply to/mention/retweet these organizations. The Twitter activities were collected in September 2015 using the GNIP firehose. Uberlink then processed the activity data to produce network and text datasets, which we provide here.

Uberlink provides the following:

- **Raw data**: network data (in graphml format), and text content data (in JSON format). Note that these data are not what Uberlink collects from GNIP: they have been processed by Uberlink, but they are in a file format that is designed for efficient storage of complex and potentially large-scale data. The current version of the raw data file format is **version 1.0**.
- **Rehydrated data**: these data files have been extracted from (or "rehydrated") the raw data files and they are in a format that is ready for analysis via third-party software.
- **Perl rehydration scripts**: these scripts are provided so users can see exactly how we produce the rehydrated data files. These scripts are released under the MIT License (`https://choosealicense.com/licenses/mit/`) and users are free to modify and share them.
- **R example script**: this script demonstrates how to create a network in R/igraph.

1 Conditions of Use of Uberlink Twitter Analytics Service

Users who purchase a TAS dataset will be required to agree to terms and conditions. These include the following terms and conditions that also apply to the use of the TAS demonstration dataset:

Uberlink cannot be held responsible for any offensive language that may be contained in Tweet content ("payload").

NOTE THAT UBERLINK IS NOT A REPOSITORY, RESELLER OR RESYNDICATOR OF TWEET CONTENT AND ANY ATTEMPT TO REVERSE ENGINEER NETWORK INFORMATION FOR THE PURPOSES OF STORING, RESELLING OR RESYNDICATING TWEET CONTENT IS EXPRESSELY PROHIBITED. FOR FURTHER INFORMATION SEE E.G. THE TWITTER TERMS

*Uberlink Corp & The Australian National University; `rob@uberlink.com`
†Uberlink Corp; `jamsheed@uberlink.com`

OF SERVICE (http://twitter.com/tos) AND DEVELOPER POLICY (https://dev.twitter.com/overview/terms/agreement-and-policy).

2 Raw data: graphml network and text content data

These are stored in the folder data_from_uberlink/

2.1 Graphml network data

The raw graphml network file contains a multiplex network.

The following examples come from the file twitter_multinet_test_period1.graphml.gz.

Twitter activities are stored as JSON strings which are graph attributes. The following is an example tweet:

```
<data key="g_641792693182402560">{&quot;content&quot;: &quot;Thank you Australia. You
    made this happen. Our members are now putting plans in place to ensure smooth
    resettlement http://t.co/ojt5qXeQnu&quot;, &quot;postedTime&quot;: &quot;2015-09-10
    T01:57:56.000Z&quot;, &quot;retweets&quot;: [[&quot;641794034600570880&quot;, &quot
    ;2015-09-10T02:03:16.000Z&quot;], [&quot;641795489713647616&quot;, &quot;2015-09-10
    T02:09:03.000Z&quot;], [&quot;641795615358218240&quot;, &quot;2015-09-10T02
    :09:33.000Z&quot;], [&quot;641793355240726528&quot;, &quot;2015-09-10T02:00:34.000Z&
    quot;], [&quot;641800234495004672&quot;, &quot;2015-09-10T02:27:54.000Z&quot;]]}</
    data>
```

The tweet id (641792693182402560) is contained in the attribute key. In addition to the payload (content) and time when posted, there is an array containing information on the retweets of this tweet (id and post time of the retweet). This example tweet was retweeted 5 times, and for each retweet there is the retweet id, and the time of the retweet.

The following are two example network nodes:

```
<node id="n176">
  <data key="v_profile">[{&quot;collectionDate&quot;: &quot;2015-09-10T01:57:56.000Z&
      quot;, &quot;data&quot;: {&quot;displayName&quot;: &quot;Refugee Council&quot;, &
      quot;utcOffset&quot;: &quot;-36000&quot;, &quot;statusesCount&quot;: 5109, &quot;
      link&quot;: &quot;http://www.twitter.com/OzRefugeeCounc&quot;, &quot;image&quot;:
      &quot;https://pbs.twimg.com/profile_images/2533603264/0el0u5lorzuwkg82mrqj_normal.
      jpeg&quot;, &quot;twitterTimeZone&quot;: &quot;Hawaii&quot;, &quot;summary&quot;:
      &quot;The Refugee Council of Australia is the national body representing 200
      organisations &amp; thousands of individuals who work with and for refugees &amp;
      asylum seekers.&quot;, &quot;postedTime&quot;: &quot;2011-06-08T03:42:54.000Z&quot
      ;, &quot;listedCount&quot;: 156, &quot;objectType&quot;: &quot;person&quot;, &quot
      ;id&quot;: &quot;id:twitter.com:313083365&quot;, &quot;favoritesCount&quot;: 242,
      &quot;preferredUsername&quot;: &quot;OzRefugeeCounc&quot;, &quot;links&quot;: [{&
      quot;rel&quot;: &quot;me&quot;, &quot;href&quot;: &quot;http://www.refugeecouncil.
      org.au&quot;}], &quot;friendsCount&quot;: 1114, &quot;followersCount&quot;: 7776,
      &quot;location&quot;: {&quot;displayName&quot;: &quot;Australia&quot;, &quot;
      objectType&quot;: &quot;place&quot;}, &quot;languages&quot;: [&quot;en&quot;], &
      quot;verified&quot;: false}}]</data>
  <data key="v_reply">{&quot;degree&quot;: 5}</data>
  <data key="v_mention">{&quot;degree&quot;: 14}</data>
  <data key="v_retweet">{&quot;degree&quot;: 7}</data>
</node>

<node id="n3046">
```

```
  <data key="v_profile">[{&quot;collectionDate&quot;: &quot;2015-09-10T02:00:34.000Z&
      quot;, &quot;data&quot;: {&quot;displayName&quot;: &quot;Matt Ross &quot;, &quot;
      utcOffset&quot;: &quot;36000&quot;, &quot;statusesCount&quot;: 172725, &quot;link&
      quot;: &quot;http://www.twitter.com/Matt_Ros&quot;, &quot;image&quot;: &quot;https
      ://pbs.twimg.com/profile_images/1424005045/image_normal.jpg&quot;, &quot;
      twitterTimeZone&quot;: &quot;Sydney&quot;, &quot;summary&quot;: &quot;Copy Editor:
       \u201cKnock knock.\u201d Photo Editor: \u201cWho\u2019s there?\u201d Copy Editor:
       \u201cTo.\u201d Photo Editor: \u201cTo who?\u201d Copy Editor: \u201cTo WHOM!\
      u201d&quot;, &quot;postedTime&quot;: &quot;2009-10-11T01:23:19.000Z&quot;, &quot;
      listedCount&quot;: 90, &quot;objectType&quot;: &quot;person&quot;, &quot;id&quot;:
       &quot;id:twitter.com:81484701&quot;, &quot;favoritesCount&quot;: 21414, &quot;
      preferredUsername&quot;: &quot;Matt_Ros&quot;, &quot;links&quot;: [{&quot;rel&quot
      ;: &quot;me&quot;, &quot;href&quot;: null}], &quot;friendsCount&quot;: 2038, &quot
      ;followersCount&quot;: 1999, &quot;location&quot;: {&quot;displayName&quot;: &quot
      ;NSW Central Coast&quot;, &quot;objectType&quot;: &quot;place&quot;}, &quot;
      languages&quot;: [&quot;en&quot;], &quot;verified&quot;: false}}, {&quot;
      collectionDate&quot;: &quot;2015-09-10T01:25:06.000Z&quot;, &quot;data&quot;: {&
      quot;statusesCount&quot;: 172716, &quot;favoritesCount&quot;: 21413}}]</data>
  <data key="v_reply">{&quot;degree&quot;: 2}</data>
  <data key="v_mention">{&quot;degree&quot;: 4}</data>
  <data key="v_retweet">{&quot;degree&quot;: 1}</data>
</node>
```

The node id is an internal sequential id number.

The "v_profile" attribute contains a JSON string with snapshots from the user's Twitter profile. The first entry in the array is a complete snapshot of the profile as at a particlar date (collectionDate) during the period. See `http://support.gnip.com/sources/twitter/data_format.html` for further information on the profile fields. Generally, this will be the oldest date (during the period) for which we have collected profile data for this user. Thereafter, the entries in the array show the profile attributes that were different (compared to the complete snapshot) on particular dates.

The "v_reply", "v_mention" and "v_retweet" attributes are degree centrality for this actor in the respective networks.

The following is an example of a network edge:

```
<edge source="n3046" target="n176">
  <data key="e_content">{&quot;retweet&quot;: [{&quot;originalID&quot;: &quot
      ;641792693182402560&quot;, &quot;retweetID&quot;: &quot;641793355240726528&quot;,
      &quot;retweetDate&quot;: &quot;2015-09-10T02:00:34.000Z&quot;}], &quot;mention&
      quot;: [{&quot;originalID&quot;: &quot;641792693182402560&quot;}]}</data>
  <data key="e_weight">{&quot;retweet&quot;: 1, &quot;mention&quot;: 1}</data>
</edge>
```

This shows that there is a directed edge from actor with (internal) id n3046 to actor n176. Actor n3046 authored an activity with id 641793355240726528: this is a retweet of the tweet with id 641792693182402560, and the retweet time was 2015-09-10T02:00:34.000Z. This activity results in a retweet edge from n3046 to n176 (the author of the original tweet). By construction, all retweet edges are also mention edges, and so the mention edge is also listed.

2.2 Text content data

The text content data files contain all of the activities that have been collected for the period. The following JSON record is from the example file twitter_text_test_period1.json.gz.

```
{"user_handle": "awfcomau", "hashtags": ["ruok", "ruokday"], "type": "retweet", "user_id":
    "2202839022", "text": "RT @ruokday: Did someone check in when you needed it most? Thank
```

them today: http://t.co/cS5uRAyNg4 #ruok #ruokday http://t.co/9NFO3eLWGG", "timestamp": "2015-09-10T00:30:59.314000Z", "retweeted_id": "641618211268509697", "tweet_id": "641770755001200640"}

This activity was authored by Twitter user @awfcomau, with Twitter user id 2202839022. The activity payload (text) and the time it was authored (timestamp) is also shown, as are the hashtags extracted from the payload. This activity has a Twitter id of 641770755001200640 (tweet_id), but it was in fact a retweet of another tweet with id 641618211268509697 (retweeted_id).

3 Rehydrated data

These are stored in the folder `pajek/`

As noted above, the graphml files are complex since they contain dynamic node and edge data and are designed for efficient storage. We are also providing network data that has been "rehydrated" from the graphml, and is in a format that ready for use in third-party tools such as R/igraph. In addition to providing data that can be easily imported into R/igraph and other software, the rehydration process achieves several other purposes:

- The graphml files are generated for a given period (typically, a month). The start/end dates of collection will not typically match the start/end of a month (they will reflect when the customer requested the data collection). The rehydration process allows the construction of networks defined for particular periods e.g. monthly, weekly, daily.
- The graphml files use network node id numbers that are sequential within a given period, and these id numbers are re-used across periods. So the node n1 in period 1 and the node n1 in period 2 are unlikely to be the same Twitter user. In contrast, the rehydrated networks use Twitter usernames/handles as the node identifier.
- As noted above, the graphml files provide dynamic snapshots of profile data e.g. the number of followers of a particular user are reported at different time points. The rehydration code produces networks for given time periods (monthly, weekly, daily) with the profile data reflecting the period. For example, for networks constructed on a weekly basis, the node attribute file will report the maximum number of followers the user had for each week.

In the rehydration code, the customer can select to produce the following types of networks: retweet, reply, mention, self-loop, and multi (multiplex i.e. all four edge types). The "self-loop" edge type is used so that target tweets that do not result in a network tie (retweet, reply, mention) can be included in the network.

So, in the demonstration dataset, there is a self-loop edge resulting from tweet with id 641765092896325633, authored by @AACTA at time 2015-09-10T00:08:30.558000Z, with payload "GIVEAWAY — Win tix to LIFE, starring #RobertPattinson, Dane DeHaan and Joel Edgerton: https://t.co/GKwPausuYB http://t.co/DTen5SpX92". This is a target activity, because the author of the tweet is one of the advocacy groups we were collecting tweets authored by. But the tweet does not result in any network ties. So for completeness, we include this activity in the network via the creation of a self-loop edge from @AACTA to itself.

The rehydrated data are in Pajek network format (with several data files per network). As noted above, the customer can choose to generate one network per edge type (reply, mention, retweet, self-loop, multi), per period (month, week, day).

The naming convention for the network files is as follows:

[study_tag].[edge_type].[included_actors].[period].[date].[file_type]

where:

- study_tag - "test", for the demonstration dataset
- edge_type retweet, mention, reply, self-loop, multi
- included_actors - "all"
- period - monthly/weekly/daily
- date timestamp in YYYY-MM/YYYY-WW/YYYY-MM-DD format
- file_type:
  - net – pajek format network dataset
  - vertex_attributes.csv – vertex attributes:
    - nodeId – internal node index
    - preferredUsername2 – Twitter username/handle
    - followersCount – number of followers (from Twitter profile)
    - friendsCount – number of 'friends' (users followed by this user) (from Twitter profile)
    - statusesCount – number of tweets/retweets ever authored by this user (from Twitter profile)
  - edge_attributes.csv – edge attributes:
    - edgeId – internal edge index
    - tweetId – Twitter id number of the activity that led to this network tie
    - edgeType – retweet/mention/reply/self-loop/multi
    - postedTime – timestamp of creation of this activity
  - payload.json – Tweet/retweet 'payloads' (content), with fields:
    - content – tweet payload
    - tweetId – Twitter id number of the activity

Note that in order to facilitate temporal comparison, all networks contain the same actors.

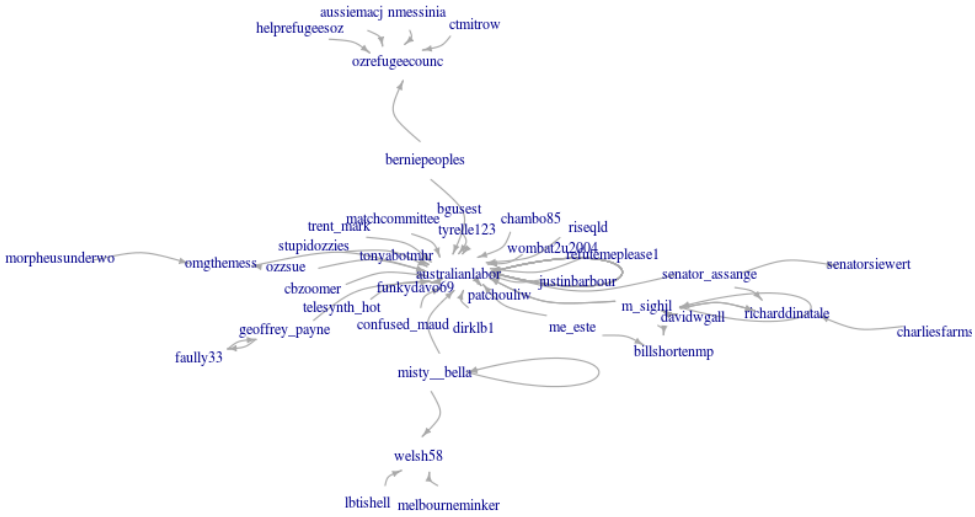
4 Perl rehydration scripts

The main purpose of the rehydration scripts is to provide insight into how to extract "research-ready" network datasets from the graphml. The rehydration scripts are written in Perl and are released under the MIT License (`https://choosealicense.com/licenses/mit/`); TAS users are welcome to modify and improve the scripts (please let us know about bugs and improvements).

5 R example script

The R script "test_network.R" demonstrates how to create in R/igraph a network for a particular period and edge type, including the assigning of vertex and edge attributes.

The following network map is constructed from the "test.multi.all.daily.2015-09-10" data files; it is the giant component when only reply edges are included.


aussiemacj
nmessinia
helprefugeesoz
ctmitrow
ozrefugeecounc
berniepeoples
bgusest
matchcommittee
chambo85
trent_mark
tyrelle123
riseqld
stupidozzies
wombat2u2004
morpheusunderwo
omgthemess
tonyabotmhr
refutemeplease1
senatorsiewert
ozzsue
australianlabor
senator_assange
cbzoomer
funkydavo69
justinbarbour
patchoulliw
telesynth_hot
m_sighil
davidwgall
richarddinatale
geoffrey_payne
confused_maud
dirklb1
me_este
charliesfarms
billshortenmp
faully33
misty__bella
welsh58
lbtishell
melbourneminker