# Amplero / University of Washington

# MSDS Capstone Project – Multivariate Time Series Analysis

# Dataset Description

## Summary

- The dataset consists of usage statistics and social attributes time series for approximately 12M prepaid mobile phone users. A random sample of 1M users were selected. Subsequently, the users belonging to the ego networks of these 1M users were added. Finally, the users belonging to the ego networks of these additional users were also added, for a total of 12 M users.
- Each user is identified by a unique entityId. The user ids and other information have been anonymized. Each user is also known as an entity.
- To each entityId correspond many time series. The period spanned by the time series is at least 120 days, ending Sept. 27, 2017.
- Usage time series capture the voice, sms, data usage of each entity. Examples: VoiceCallsPerDayTimeSeries, CarrierReportedSubscriptionStateDeltaTimeSeries.
- Social time series capture the social network structure of each entity. Examples: OutboundSMSNetworkPageRankLast7DaysTimeSeries, InboundSMSNetworkOnNetFractionOfDegreeLast7DaysTimeSeries.
- State and event time series capture the status of an entity and discrete events relevant to that entity. Examples: RechargeTimeSeries, PlanDeltaTimeSeries, CallsToCareAgentTimeSeries.

## File Structure

Data are in 51 gzipped files. There are three types of files: a) 10 usage*.jsonish.gz files contain the usage data; b) 10 social*.jsonish.gz files contain social network data; c) one file called state_recharge_etc.jsonish.gz contains state, event, and some usage data.

All files are in the JSON format. As an example, the figure below shows the beginning of **state_recharge_etc.jsonish.gz**. Notice the entityId at the top, for which we can see 2 attributes; each attribute is a time series. For this entityId there were no calls to customer care, but several phone calls per day were made.

```
{
  "entityId": "d008919a21760f60b8c7fbc338921153",
  "attributes": {
    "CallsToCareAgentPerDayTimeSeries": {
      "value": []
    },
    "VoiceCallsPerDayTimeSeries": {
      "value": [
        {
          "value": 21,
          "timestamp": 1491091200000
        },
        {
          "value": 51,
          "timestamp": 1491177600000
        },
        {
          "value": 50,
          "timestamp": 1.491264e+12
        },
        {
          "value": 35,
          "timestamp": 1491350400000
        },
        {
          "value": 64,
          "timestamp": 1491436800000
        },
```


*Figure 1: Example time series in JSON format.*

For each entityId, the following time series are contained in **state_recharge_etc.jsonish.gz.** The ones that are particularly important are  starred, with a short explanation.

"CallsToCareAgentPerDayTimeSeries": {
"VoiceCallsPerDayTimeSeries": { ** number of calls per day
"RichCarrierSubscriptionStateDeltaTimeSeries": {
"CallsToCanadaPerDayTimeSeries": {
"CreditAppliedTimeSeries": {
"AutoPayDeltaTimeSeries": {
"CarrierReportedSubscriptionStateDeltaTimeSeries": { ** is an entityID active or inactive? Needed for churn
"RechargeTimeSeries": { ** amount and time of a pre-paid account recharge
"CarrierRechargeTimeSeries": {
"BonusAmountTimeSeries": {
"FamilyPlanTimeSeries": {
"ThrottledDeltaTimeSeries": {
"CallsToInternationalCallingCardPerDayTimeSeries": {
"DroppedCallsPerDayCountTimeSeries": {
"DeviceDeltaTimeSeries": {
"CompetitorWebsiteVisitedTimeSeries": {
"InferredSuscriptionStateDeltaOutboundUsageGrace30TimeSeries": {
"BlockedCallsPerDayCountTimeSeries": {

"PortOutTimeSeries": { ** has entity ported out of the carrier?
"PortInTimeSeries": {
"InferredSuscriptionStateDeltaOutboundUsageAndRechargeGrace30TimeSeries": {
"DroppedAndBlockedCallsPerDayCountTimeSeries": {
"FamilyPlanRechargeTimeSeries": {
"PlanDeltaTimeSeries": {
"CallsToCustomerCarePerDayTimeSeries": {
"CallsToMexicoPerDayTimeSeries": {
"InferredSuscriptionStateDeltaRechargeGrace30TimeSeries":

Time series are sparse, meaning that a (timestamp, value) is present only if there is something to report for that value. For usage time series (e.g. VoiceCallsPerDayTimeSeries) this means that missing values are implied to be 0; for state time series (e.g. the various *SubscriptionStateTimeSeries) the missing values are implied to be identical to the last known value, until the value changes (so these are really piece-wise constant).

The **usage*.jsonish.gz** files contain usage information for each entity. The usage is according to SMS, Data, and Voice. These are the relevant time series:

"CompactSMSPerDayTimeSeries": {
"CompactDataKBPerDayTimeSeries": {
"CompactVoiceSecondsPerDayTimeSeries": {

However, these time series are nested. The idea is that SMS usage time series can be sub-divided in component time series according to a finer-grained classification of usage: inbound SMS, outbound SMS, outbound SMS to international recipients, inbound SMS from international senders, etc. and similarly for other dimensions of usage. Here is an example:

"CompactSMSPerDayTimeSeries": {
"OutboundNotInNetworkInternationalCompactSMSPerDayTimeSeries": [
"OutboundInNetworkNotInternationalCompactSMSPerDayTimeSeries": [
"NotInNetworkInternationalCompactSMSPerDayTimeSeries": [
"NotOutboundNotInNetworkCompactSMSPerDayTimeSeries": [
"NotOutboundNotInternationalCompactSMSPerDayTimeSeries": [
"OutboundInNetworkCompactSMSPerDayTimeSeries": [
"NotOutboundNotInNetworkNotInternationalCompactSMSPerDayTimeSeries": [
"OutboundCompactSMSPerDayTimeSeries": [
"NotInNetworkNotInternationalCompactSMSPerDayTimeSeries": [
"NotOutboundNotInNetworkInternationalCompactSMSPerDayTimeSeries": [
"InNetworkCompactSMSPerDayTimeSeries": [
"OutboundNotInternationalCompactSMSPerDayTimeSeries": [
"InternationalCompactSMSPerDayTimeSeries": [
"NotOutboundCompactSMSPerDayTimeSeries": [
"OutboundInternationalCompactSMSPerDayTimeSeries": [
"OutboundNotInNetworkCompactSMSPerDayTimeSeries": [

"InNetworkNotInternationalCompactSMSPerDayTimeSeries": [
"NotOutboundInNetworkNotInternationalCompactSMSPerDayTimeSeries": [
"NotOutboundInternationalCompactSMSPerDayTimeSeries": [
"NotInNetworkCompactSMSPerDayTimeSeries": [
"NotOutboundInNetworkCompactSMSPerDayTimeSeries": [

The social*.jsonish.gz files contain time series of social network attributes for each entity. The social networks are constructed from the voice and SMS traffic, and the members of each of these social networks may not be the same for a given entity (a user may never call someone but may text that person frequently).

In social network analysis, Degree means number of connections (which is equal to the number of members in the ego network). Degree can be broken out further by inbound/outbound, in-network/out-of-network. Below is a list of the many time series available for social network behavior.

"InboundSMSNetworkDegreeLast14DaysTimeSeries": {
"InboundSMSNetworkDegreeLast28DaysTimeSeries": {
"InboundSMSNetworkDegreeLast7DaysTimeSeries": {
"InboundSMSNetworkOffNetDegreeLast14DaysTimeSeries": {
"InboundSMSNetworkOffNetDegreeLast28DaysTimeSeries": {
"InboundSMSNetworkOffNetDegreeLast7DaysTimeSeries": {
"InboundSMSNetworkOffNetFractionOfDegreeLast14DaysTimeSeries": {
"InboundSMSNetworkOffNetFractionOfDegreeLast28DaysTimeSeries": {
"InboundSMSNetworkOffNetFractionOfDegreeLast7DaysTimeSeries": {
"InboundSMSNetworkOnNetDegreeLast14DaysTimeSeries": {
"InboundSMSNetworkOnNetDegreeLast28DaysTimeSeries": {
"InboundSMSNetworkOnNetDegreeLast7DaysTimeSeries": {
"InboundSMSNetworkOnNetFractionOfDegreeLast14DaysTimeSeries": {
"InboundSMSNetworkOnNetFractionOfDegreeLast28DaysTimeSeries": {
"InboundSMSNetworkOnNetFractionOfDegreeLast7DaysTimeSeries": {
"InboundVoiceCountNetworkDegreeLast14DaysTimeSeries": {
"InboundVoiceCountNetworkDegreeLast28DaysTimeSeries": {
"InboundVoiceCountNetworkDegreeLast7DaysTimeSeries": {
"InboundVoiceCountNetworkOffNetDegreeLast14DaysTimeSeries": {
"InboundVoiceCountNetworkOffNetDegreeLast28DaysTimeSeries": {
"InboundVoiceCountNetworkOffNetDegreeLast7DaysTimeSeries": {
"InboundVoiceCountNetworkOffNetFractionOfDegreeLast14DaysTimeSeries": {
"InboundVoiceCountNetworkOffNetFractionOfDegreeLast28DaysTimeSeries": {
"InboundVoiceCountNetworkOffNetFractionOfDegreeLast7DaysTimeSeries": {
"InboundVoiceCountNetworkOnNetDegreeLast14DaysTimeSeries": {
"InboundVoiceCountNetworkOnNetDegreeLast28DaysTimeSeries": {
"InboundVoiceCountNetworkOnNetDegreeLast7DaysTimeSeries": {

"InboundVoiceCountNetworkOnNetFractionOfDegreeLast14DaysTimeSeries": {
"InboundVoiceCountNetworkOnNetFractionOfDegreeLast28DaysTimeSeries": {
"InboundVoiceCountNetworkOnNetFractionOfDegreeLast7DaysTimeSeries": {
"OutboundSMSNetworkDegreeLast14DaysTimeSeries": {
"OutboundSMSNetworkDegreeLast28DaysTimeSeries": {
"OutboundSMSNetworkDegreeLast7DaysTimeSeries": {
"OutboundSMSNetworkOffNetDegreeLast14DaysTimeSeries": {
"OutboundSMSNetworkOffNetDegreeLast28DaysTimeSeries": {
"OutboundSMSNetworkOffNetDegreeLast7DaysTimeSeries": {
"OutboundSMSNetworkOffNetFractionOfDegreeLast14DaysTimeSeries": {
"OutboundSMSNetworkOffNetFractionOfDegreeLast28DaysTimeSeries": {
"OutboundSMSNetworkOffNetFractionOfDegreeLast7DaysTimeSeries": {
"OutboundSMSNetworkOnNetDegreeLast14DaysTimeSeries": {
"OutboundSMSNetworkOnNetDegreeLast28DaysTimeSeries": {
"OutboundSMSNetworkOnNetDegreeLast7DaysTimeSeries": {
"OutboundSMSNetworkOnNetFractionOfDegreeLast14DaysTimeSeries": {
"OutboundSMSNetworkOnNetFractionOfDegreeLast28DaysTimeSeries": {
"OutboundSMSNetworkOnNetFractionOfDegreeLast7DaysTimeSeries": {
"OutboundSMSNetworkPageRankLast7DaysTimeSeries": {
"OutboundSMSNetworkPageRankQuantileLast7DaysTimeSeries": {
"OutboundVoiceCountNetworkDegreeLast14DaysTimeSeries": {
"OutboundVoiceCountNetworkDegreeLast28DaysTimeSeries": {
"OutboundVoiceCountNetworkDegreeLast7DaysTimeSeries": {
"OutboundVoiceCountNetworkOffNetDegreeLast14DaysTimeSeries": {
"OutboundVoiceCountNetworkOffNetDegreeLast28DaysTimeSeries": {
"OutboundVoiceCountNetworkOffNetDegreeLast7DaysTimeSeries": {
"OutboundVoiceCountNetworkOffNetFractionOfDegreeLast14DaysTimeSeries": {
"OutboundVoiceCountNetworkOffNetFractionOfDegreeLast28DaysTimeSeries": {
"OutboundVoiceCountNetworkOffNetFractionOfDegreeLast7DaysTimeSeries": {
"OutboundVoiceCountNetworkOnNetDegreeLast14DaysTimeSeries": {
"OutboundVoiceCountNetworkOnNetDegreeLast28DaysTimeSeries": {
"OutboundVoiceCountNetworkOnNetDegreeLast7DaysTimeSeries": {
"OutboundVoiceCountNetworkOnNetFractionOfDegreeLast14DaysTimeSeries": {
"OutboundVoiceCountNetworkOnNetFractionOfDegreeLast28DaysTimeSeries": {
"OutboundVoiceCountNetworkOnNetFractionOfDegreeLast7DaysTimeSeries": {
"OutboundVoiceCountNetworkPageRankLast7DaysTimeSeries": {
"OutboundVoiceCountNetworkPageRankQuantileLast7DaysTimeSeries": {

## Tips

The total size of the files is about 400GB compressed. It will be large once uncompressed. But you may not need to do that: software packages exist that allow you to import and explore the data while leaving the original files as they are. The Python gzip package is an example.

The command-line utility jq is helpful for formatting JSON files on the screen, on linux. Say you want to inspect a gzipped JSON file; do this:
zcat social1.jsonish.gz | jq . | less

A very useful Python package is the json() package: you will need it to read the time series from file.

There are many time series to consider: start with a few, but diverse (covering different types of characteristics), and build on that. For nested time series, start with the top-level ones, or with the coarsest level of inbound/outbound series.

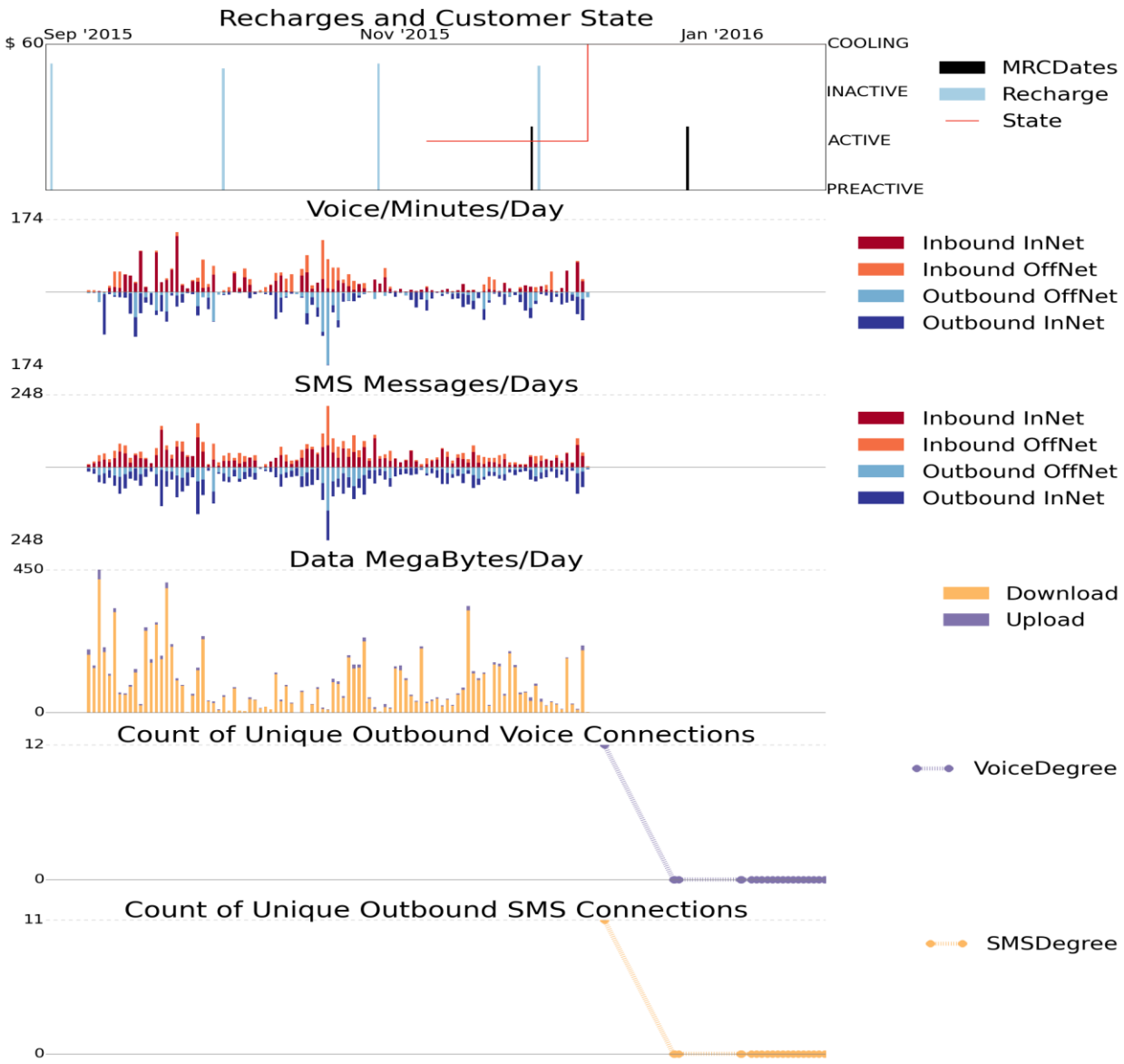


*Figure 2: Example of multivariate time series that can be leveraged for predicting future user behavior for a prepaid mobile use case. The provided data set may differ.*