SocialMediaLab Tutorial - ICA2017

Tim Graham and Rob Ackland

25 May 2017

Introduction

This is a tutorial for the *SocialMediaLab* R package. In this tutorial you will learn how to collect social media data from Twitter (using the #ica17 hashtag), create networks, and perform basic social network analysis (SNA).

SocialMediaLab enables users to collect social media data and create different kinds of networks for analysis. It is a 'Swiss army knife' for this kind research, enabling a swift work flow from concept to data to fully-fledged network, ready for SNA and other analysis. Drawing on several key R packages, it can handle large datasets and create very large networks, upwards of a million or more nodes (depending on your computer's resources). The following data sources are currently supported, although in this tutorial we will only be collecting data from Twitter:

1. Facebook
2. YouTube
3. Twitter
4. Instagram (although API access is extremely limited without an approved app)

Installation and setup

First ensure that the *SocialMediaLab* package is installed and loaded.

We also want to install the *magrittr* package, so we can simplify the work flow by using 'verb' functions that pipe together. We will also be using the *igraph* package for network analysis.

The following commands will check if the packages are installed and install them as necessary, then load them.

Note: SocialMediaLab is available as an official package on CRAN, but the latest development version is available on GitHub. We suggest installing the latest version from GitHub, using the code below.

Note: **Recent changes in the `httr` package caused problems for the `twitteR` package. We resolve this using a quick-fix by installing an earlier version of `httr`. However, first we have to install the most recent version of `httr` package, before downgrading it to the earlier version.**

```
install.packages("httr")
if (!"devtools" %in% installed.packages()) install.packages("devtools")
require(devtools)
devtools::install_version("httr", version="0.6.0", repos="http://cran.us.r-project.org")

if (!"SocialMediaLab" %in% installed.packages()) {
  devtools::install_github("voson-lab/SocialMediaLab/SocialMediaLab")
}
require(SocialMediaLab)

if (!"magrittr" %in% installed.packages()) install.packages("magrittr")
require(magrittr)
```

```
if (!"igraph" %in% installed.packages()) install.packages("igraph")
require(igraph)
```

```
## Loading required package: SocialMediaLab
```

```
## Loading required package: magrittr
```

```
## Loading required package: igraph
```

```
##
## Attaching package: 'igraph'
```

```
## The following object is masked from 'package:magrittr':
##
##     %>%
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
```

```
## The following object is masked from 'package:base':
##
##     union
```

You will also need to get API access for Twitter. You will *not* be able to collect any data until you have acquired API credentials. Step-by-step instructions for obtaining API access are available from the VOSON website.

Twitter data collection and analysis

In this section we will run through how to collect data from Twitter, create networks, and perform different kinds of analysis.

It is currently possible to create 3 different types of networks using Twitter data collected with `SocialMediaLab`. These are (1) *actor* networks; (2) *bimodal* networks; and (3) *semantic* networks. In this session we will create an *actor* and a *semantic* network (we created a bimodal Facebook network in the previous section).

First, define the API credentials. Due to the Twitter API specifications, it is not possible to save authentication token between sessions. The `Authenticate()` function is called only for its side effect, which provides access to the Twitter API for the current session.

```
# REPLACE WITH YOUR API KEY
myapikey <- "xxxx"
# REPLACE WITH YOUR API SECRET
myapisecret <- "xxxx"
# REPLACE WITH YOUR ACCESS TOKEN
myaccesstoken <- "xxxx"
# REPLACE WITH YOUR ACCESS TOKEN SECRET
myaccesstokensecret <- "xxxx"
```

Given that we are going to be creating two different types of Twitter networks (actor and semantic), we will `Collect()` the data, but not pipe it directly through to `Network()` straight away. This means we can reuse the data multiple times to create two different kinds of networks for analysis.

We will collect 150 recent tweets that have used the #ica17 hashtag. This is the dominant hashtag for Australian politics.

```
myTwitterData <- Authenticate("twitter",
                              apiKey=myapikey,
```

```
                        apiSecret=myapisecret,
                        accessToken=myaccesstoken,
                        accessTokenSecret=myaccesstokensecret) %>%
                Collect(searchTerm="#ica17",
                        numTweets=500,
                        writeToFile=FALSE,
                        verbose=TRUE)
```

```
## [1] "Using direct authentication"
## Now retrieving data based on search term: #ica17
## Done
## Cleaning and sorting the data...
## Done
```

We can have a quick look at the data we just collected:

```
View(myTwitterData)
```

Note the class of the dataframe, which lets `SocialMediaLab` know that this is an object of class `dataSource`, which we can then pass to the `Create()` function to generate different kinds of networks:

```
class(myTwitterData)
```

```
## [1] "data.frame" "dataSource" "twitter"
```

First, we will create an *actor* network. In this actor network, edges represent interactions between Twitter users. An interaction is defined as a 'mention' or 'reply' or 'retweet' from user i to user j, given 'tweet' m. In a nutshell, a Twitter actor network shows us who is interacting with who in relation to a particular hashtag or search term.

```
g_twitter_actor <- myTwitterData %>% Create("Actor")
```

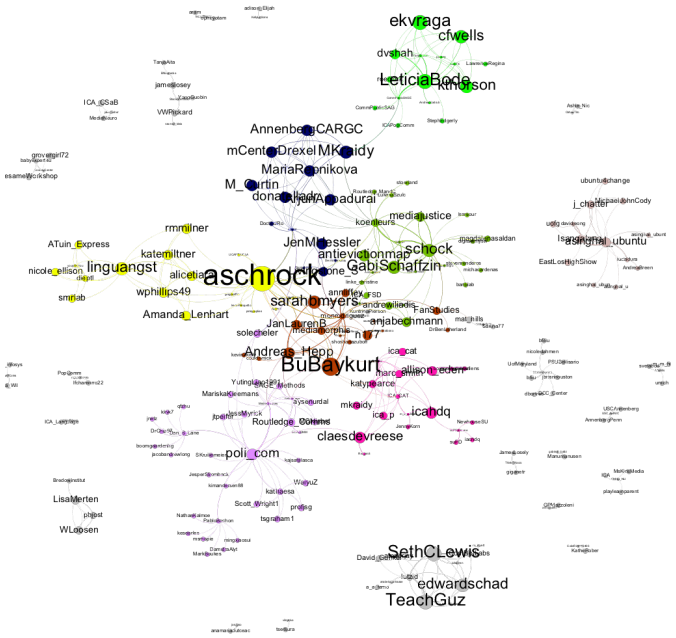
```
## Generating the network...
##
## Done.
```

We can now examine the description of our network:

```
g_twitter_actor
```

```
## IGRAPH DN-- 286 780 --
## + attr: name (v/c), label (v/c), edgeType (e/c), timeStamp (e/c),
## | tweet_id (e/c)
## + edges (vertex names):
##  [1] CIMA_Media     ->susan_abbott    a_a_tamo       ->TeachGuz
##  [3] kattyalhayek   ->icahdq          DrBillASU      ->JanLaurenB
##  [5] johne326       ->LboroSocSci     karen__harper  ->MeasureRadio
##  [7] karen__harper  ->linke_christine nathansources  ->alicetiara
##  [9] WCSTCo         ->alicetiara      radharani_m    ->lsangalang
## [11] radharani_m    ->luciadura       radharani_m    ->lsangalang
## [13] radharani_m    ->davidjeong      RutgersCommInfo->DrSha
## + ... omitted several edges
```

Here is a visualisation of an earlier data collection of #ica17 tweets (25 May 2017, using Gephi).



Who are the top 3 important users in our #ica17 actor network? There are several ways to do this. We will use the PageRank algorithm implementation in `igraph` to calculate this:

```
pageRank_ica17_actor <- sort(page.rank(g_twitter_actor)$vector,decreasing=TRUE)
head(pageRank_ica17_actor,n=3)
```

```
## Andreas_Hepp  GPMazzoleni         kbczk
##   0.01722245   0.01519379   0.01519379
```

What about the 3 least important users (with all due respect...):

```
tail(pageRank_ica17_actor,n=3)
```

```
##   antmandan  KarinWahlJ         ehl
## 0.001930021 0.001930021 0.001930021
```

Is there any kind of community structure within the user network? As per the previous Facebook analysis we will use the infomap algorithm implementation in `igraph`.

```
imc <- infomap.community(g_twitter_actor, nb.trials = 1) # increase nb.trials for better quality commun

# create a vector of users with their assigned community number
communityMembership_ica17 <- membership(imc)
# summarise the distribution of users to communities
commDistribution <- summary(as.factor(communityMembership_ica17))
# which community has the max number of users
tail(sort(commDistribution),n=1)
```

```
##   1
## 143
```

```
# create a list of communities that includes the users assigned to each community
communities_ica17 <- communities(imc)
# look at the members of the most populated community
communities_ica17[names(tail(sort(commDistribution),n=1))]
```

```
## $`1`
```

```
##   [1] "DrBillASU"       "radharani_m"     "RutgersCommInfo"
##   [4] "Bsheridan2208"   "freddy_hopp"     "MaryChayko"
##   [7] "balicea1"        "MsDianaLee"      "omkouture"
##  [10] "allison_eden"    "aschrock"        "dani_isu"
##  [13] "JeannineRelly"   "schock"          "ashwinnag"
##  [16] "fromanelli41"    "j_chatter"       "simone_natale"
##  [19] "ICAGames"        "BrettOppegaard"  "saidtuzel"
##  [22] "iambrandao"      "richardhuskey"   "MediaNeuro"
##  [25] "antievictionmap" "jmgrygiel"       "lutzid"
##  [28] "StewartColes"    "ciullalipkin"    "Crafty_AI"
##  [31] "nickkauf"        "Jugdev"          "MediaLitBot"
##  [34] "DevinaSarwatay"  "filippotrevisan" "BostonJoan"
##  [37] "nodexl"          "AncaMatioc"      "arbitist"
##  [40] "Dan_S_Lane"      "SebaValenz"      "poli_com"
##  [43] "reneehobbs"      "lsangalang"      "bradleyjbond"
##  [46] "Mediatingmimi"   "davidjeong"      "Livingstone_S"
##  [49] "annienavar"      "gabriele_balbi"  "digitalamysw"
##  [52] "JanLaurenB"      "babyexpert4u"    "claesdevreese"
##  [55] "andrewiliadis"   "ehl"             "luciadura"
##  [58] "DrSha"           "JennyKorn"       "solecheler"
##  [61] "AnnenbergPenn"   "ICA_CSaB"        "ICA_Language"
##  [64] "Jana_Wil"        "marc_smith"      "laiacastroh"
##  [67] "Mo_Skovsgaard"   "M_Aronczyk"      "MollyGreenwood3"
##  [70] "michelleefunk"   "lindseyblumell"  "David_Proper"
##  [73] "SMihelj"         "ubuntu4change"   "MichaelJohnCody"
##  [76] "asinghal_ubuntu" "asinghal_ubun"   "EastLosHighShow"
##  [79] "uofg"            "asinghal_ubunt"  "SDSU_JMS"
##  [82] "SDSU_Comm"       "tm_hopp"         "ValerieEBarker1"
##  [85] "ICA_PRD"         "PBSKIDS"         "SinkingShipEnt"
##  [88] "diana_ingenhoff" "dr_rjahng"       "DiMAP_UW"
##  [91] "KathyJCramer"    "mschudson2"      "shenfei1010"
##  [94] "JayeonLee"       "USCAnnenberg"    "OhioStateComm"
##  [97] "lee_nicole"      "msvandyke"       "ISU_GSJC"
## [100] "DaraMWald"       "SuJungKim_ISU"   "rcozma"
## [103] "magdalenasaldan" "onekade"         "discoursology"
## [106] "ashw"            "yonty"           "cinehead"
## [109] "LanceBennett1"   "yangyunkang"     "striphas"
## [112] "merlyna"         "EmilianoTrere"   "DataRescueSFBay"
## [115] "lborouniversity" "priydee"         "bbcmediaaction"
## [118] "marika_louise"   "laura4lano"      "AARP"
## [121] "Microsoft"       "DanielleCorple"  "jlinabary"
## [124] "mediaghosts"     "nishagarud"      "RaiAlUc"
## [127] "kkvk7"           "jnelz"           "boomgaardenhg"
## [130] "qfzhu"           "jacobandrewlong" "DrChip97"
## [133] "frankwad"        "JenMHessler"     "petervanaelst"
## [136] "jenjpan"         "mollyeroberts"   "ica_infosys"
## [139] "lsarsour"        "katypearce"      "ica_cat"
## [142] "mkraidy"         "ica_p"
```

Next, we will create a *semantic* network. In this network nodes represent unique concepts (in this case unique terms/words extracted from a set of 500 tweets), and edges represent the co-occurrence of terms for all observations in the data set. For example, for this Twitter semantic network, nodes represent either hashtags (e.g. "#ica17") or single terms ("happy"). If there are 500 tweets in the data set (i.e. 500 observations), and the hashtag *#ica17* and the term *happy* appear together in every tweet, then this would be represented by

an edge with weight equal to 500.

```
g_twitter_semantic <- myTwitterData %>% Create("Semantic")
```

```
## [1] "Generating Twitter semantic network..."
##
## Done.
```

Let's have a look at the network description:

```
g_twitter_semantic
```

```
## IGRAPH UNW- 185 598 --
## + attr: name (v/c), label (v/c), weight (e/n)
## + edges (vertex names):
##  [1] #nationalwineday  --#ica
##  [2] #nationalwineday  --#nationalwineday
##  [3] #nationalwineday  --#thursdaythoughts
##  [4] #nationalwineday  --nato
##  [5] #nationalwineday  --#rednoseday
##  [6] #nationalwineday  --#boxing
##  [7] #nationalwineday  --#whenthebossisaway
##  [8] #ica              --#starwars40th
## + ... omitted several edges
```

What are the top 10 important terms in our #ica17 semantic network? Once again we will calculate this using PageRank:

```
pageRank_ica17_semantic <- sort(page.rank(g_twitter_semantic)$vector,decreasing=TRUE)
pageRank_ica17_semantic[1:10]
```

```
##             #ica17               #ica      #mediajustice
##         0.20920798         0.09945166         0.06310954
##              media   #nationalwineday  #thursdaythoughts
##         0.02350282         0.02046237         0.02046237
##        #rednoseday             #hmc17            #boxing
##         0.01875503         0.01872424         0.01795613
## #whenthebossisaway
##         0.01795613
```

Conclusion

We hope that you enjoyed this tutorial and that you find *SocialMediaLab* useful in your research. Please feel free to contact us should you have any questions or comments (or find any errors in the document). We would love to hear your ideas and feedback.

All the best,

Tim Graham and Rob Ackland

(ICA Annual Conference, 2017)