CHAPTER 1 IntroductiontotheSocialWeb

1. Introduction to the Unit

Motivation. People are uploading data (text, images, videos, links) to the Web at every moment of the day. Some of you may be doing it right now!

With proper analysis, we can use this data to examine the state of the world and predict the future directions of the world's communities.

In this unit, we will learn how to perform this analysis.

Examining the Learning Guide. The learning guide contains a description of:

- the content of the unit
- what is expected from each student
- delivery of the unit
- the assessment

The learning guide is found in vUWS > 300958 > Unit Information

Teaching Team. Unit Convenor, Lecturer and Tutor:

- Chris D'Souza c.d'souza@city.westernsydney.edu.

2.Social Networks

What are Social Networks? Social Networks consist of a set of items and connections showing how each of the items interact.

Social networks or just networks

Which of these are considered social networks:

- a set of students, where a pair of students is connected if they attend the same class,
- a set of cities, where a pair of cities is connected if share a highway,
- a set of programming languages, where a pair is connected if they share a paradigm,
- a set of people, where a pair is connected if they are related.

Social Network Analysis is the analysis of these networks to obtain information such as which item is the most influential, and if new items are introduced how does the network change.

Small World: Six degrees of separation. The theory that two randomly selected people can find a chain of friends that is at most six in length.

To examine the small world theory an experiment was devised using the U.S. Postal Service.

- `https://en.wikipedia.org/wiki/Small_world_experiment`

Paul Erdös. The small world phenomenon extends to many communities. One of the earliest examined was the academic research community.

- The centre of the research community is thought to be Paul Erdös, who published about 1500 articles with 509 co-authors.
- Researchers can measure their centrality with respect to this research social network by computing their Erdös number.

Erdös Number. If you have an Internet connected device, use the above link to examine the Erdös Number of a few of your lecturers.

Kevin Bacon. Databases such as the IMDB have allowed us to examine how connected the set of actors are through the movies they appear in together.

It was thought that the most connected actor was Kevin Bacon.

The movie distance between any two actors can be computed at the Oracle of Bacon.

Bacon Number. If you have an Internet connected device, use the above link to examine the Bacon Number of a few actors. Can you find any actors with a Bacon number of 5 or more.

NSA Surveillance. The small world phenomenon experiment shows that we are connected to most of the world in six hops.

The NSA put anyone suspected of terrorist activity, and anyone connected to them by three or less hops, under surveilance (The Guardian).

2. The Social Web

The World Wide Web. The World Wide Web is a collection of Web pages and databases that are served from computers spread across the Internet.

The Web was conceived by Tim Berners-Lee as a method a sharing experimental particle physics results between labs across the world. It was quickly seen that this platform could be used to serve all kinds of information.

The Web is intrinsically social in that, for it to grow, it requires us to upload information that others will read.

Being Social on the Web. As time has passed, higher level applications have been created for social activity on the Web:

1. Forums: vUWS Discussion Board, Google Groups
2. Bookmarking: Delicious, Pinterest, Reddit, Slashdot
3. Video: YouTube, Vimeo
4. Images: Flickr, Instagram
5. Music: Last.fm
6. Bloging: LiveJournal, WordPress
7. Academic: Mendeley
8. Business: LinkedIn
9. Location: Foursquare
10. MySpace, Facebook, Google+
11. Microblog: Twitter, Vine, Identi.ca, Tumblr
12. Information seeking: StackExchange

Twitter. Twitter is a Web based social networking service that allows us to post messages of at most 140 characters, called tweets.

- Twitter is social in that we are able to follow the posts of others and they are able to follow us.
- The relationships at directional, meaning that A can follow B without B having to follow A.
- Everyone's tweets are public and searchable.

Twitter recently released that over 500 million tweets are posted per day, with over 1.6 billion search queries per day.

Facebook. Facebook is a general social networking service that allows users to exchange information.

- Facebook is social in that it allows us to be selective of who is able to view the information we post.
- The relationships are non-directional, meaning if A is friends with B then B is also friends with A.
- Specific information in Facebook can be restricted to selected friends and therefore not available to the public.

Facebook has over 2.2 billion users.

Contributing to the Social Web. For Web based social networks to exist, we must freely contribute and view information from them.

The recent introduction of the mobile devices (such as phones and tablet computers) that have access to the Web has added to the popularity of Web based social networks. They allow us to conveniently upload and view information as it happens.

Our presence on the Social Web: Turn to the people around you and tell them

- which Social Web services you have contributed information to.
- how often you add and read information from them
- methods which you add and view information from them
- why you add and view information from them

3. Analysis of the Social Web

Information in the Social Web. Many social networking services on the Web have become a place for us to record our lives.

- We post images of events
- We share interesting sites with others
- We write reviews of products
- We broadcast what we have has for breakfast

The social Web has become a place where we feel comfortable releasing our personal information.

Using personal information. Businesses have an interest in the Social Web:

- The top priority for all businesses is to make a profit, usually by offering a product or service.
- Businesses must invest in the development of a product or service before it can be offered to the public.
- If the public does not buy the product or service, the business had failed.

Analysis of the Social Web allows us to predict the behaviour of the public under certain situations.

Getting opinions. Governments have an interest in the Social Web:

- The role of the government is to develop and implement policies and draft laws.
- The polices and laws affect the public.
- The public decide who is voted into government.

Analysis of the Social Web allows us to gauge the opinion of the public at given times.

4. Data Analysis with R

Introducing R. This semester, we will be using R to perform our data analysis.

What is R?. R is a software environment for statistical computing and graphics. It runs on just about any platform (except iPad!) and is completely free (in the GNU sense).

It is used extensively by academic statisticians for research and teaching and is gaining ground in business.

It has 12639 extension packages available.

Pros. Its free and open source. It has most methods for most things mostly before any other package. It has the best graphics. It extendable.

Cons. It has a steep learning curve. No GUI by default. Poor (but improving) memory management; difficulty with very large data sets.

R Resources.

- `http://www.r-project.org` — Main R website.
- CRAN — `http://cran.csiro.au` — Comprehensive R Archive Network — base software and add-on packages.
- RStudio — `http://www.rstudio.com` — is a powerful IDE for R
- R Commander — `install.package(Rcmdr)` — is a partial GUI interface to R — requires TclTk.
- R Graph Gallery — `http://gallery.r-enthusiasts.com/` — loads of pretty pictures.
- `http://cran.csiro.au/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf` — "A (very) short Introduction to R"
- "Introductory Statistics with R", Peter Dalgaard, Springer 2008.

R Commands. R can be used as a basic calculator.

```
> 1+1
```

```
[1] 2
```

```
> sqrt(2)
```

```
[1] 1.414214
```

```
> 2^5
```

```
[1] 32
```

It can store things as named objects.

```
> x <- 1
> print(x)

[1] 1
```

R Commands. It understands vectors and matrices.

```
> x <- c(1,2)
> m <- matrix(c(1,2,3,4), ncol=2, byrow=TRUE)
> print(m)

     [,1] [,2]
[1,]    1    2
[2,]    3    4

> m %*% x

     [,1]
[1,]    5
[2,] 11
```

R Commands. It has functions, and you can write them.

```
> x <- sqrt(2)
> sqr <- function(x) x^2
> sqr(2)

[1] 4
```

Data in R. Tables are stored in `data.frames` .

```
> head(iris)

  Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.1 3.5    1.4    0.2

2          4.9 3.0    1.4    0.2

3          4.7 3.2    1.3    0.2

4          4.6 3.1    1.5    0.2

5          5.0 3.6    1.4    0.2

6          5.4 3.9    1.7    0.4
```

```
Species
1  setosa
2  setosa
3  setosa
4  setosa
5  setosa
6  setosa
```

Data in R.

```
> dim(iris)
[1] 150   5
```

Some columns are numeric, others are factors.

```
> sapply(iris, class)
Sepal.Length Sepal.Width Petal.Length Petal.Width
   ”numeric”    ”numeric”    ”numeric”    ”numeric”
     Species
    ”factor”
```

Data can read from text files (`read.csv` and `read.table`) and various formats using the `foreign` package.

Basic Statistics.

```
> x <- rnorm(100)
> mean(x)
[1] -0.0677875
> var(x) ### sd(x)
[1] 0.9472886
> fivenum(x)
[1] -2.94874985 -0.64575861 0.01092249 0.60585333
[5] 1.85728005
```

minimum, lower-quartile, median, upper-quartile, maximum

Basic Statistics.

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length
```

```
 Min. :4.300 Min.     :2.000 Min.    :1.000 1st
 Qu.:5.100     1st Qu.:2.800 1st Qu.:1.600
 Median :5.800 Median :3.000 Median :4.350 Mean
 :5.843 Mean   :3.057 Mean    :3.758 3rd
 Qu.:6.400     3rd Qu.:3.300 3rd Qu.:5.100 Max.
 :7.900 Max.   :4.400 Max.    :6.900
  Petal.Width  Species Min.
 :0.100 setosa :50
 1st Qu.:0.300 versicolor:50 Median
 :1.300virginica :50
 Mean :1.199 3rd
 Qu.:1.800 Max.
 :2.500
```

Basic Statistics.

```
> t.test(x)


    One Sample t-test

data: x t = -0.69648, df = 99, p-value = 0.4878
alternative hypothesis: true mean is not equal to
0 95 percent confidence interval:
 -0.2609089 0.1253339
sample estimates:
 mean of x
-0.0677875
```

R has extensive plotting.

```
> plot(Sepal.Length~Sepal.Width,col=Species,data=iris, pch=16,
+ cex.lab=0.6, cex.axis=0.6)
```

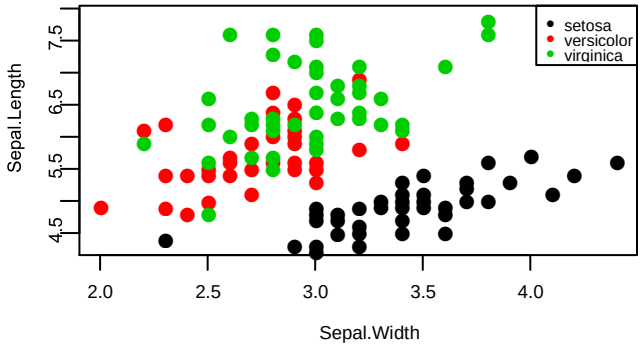

Figure 1. A scatter plot

R has extensive plotting.

```
> plot(Sepal.Length~Species, data=iris, cex.lab=0.6, cex.axis=0.6)
```

Summary.

- Social Networks describe interaction amongst a set of elements.
- Social Networks appear everywhere.
- The Web has provided us a basis for social interaction.
- Analysing social networks on the Web provides us an insight of the state of the community.
- We can use R to assist in the analysis of social Web data.

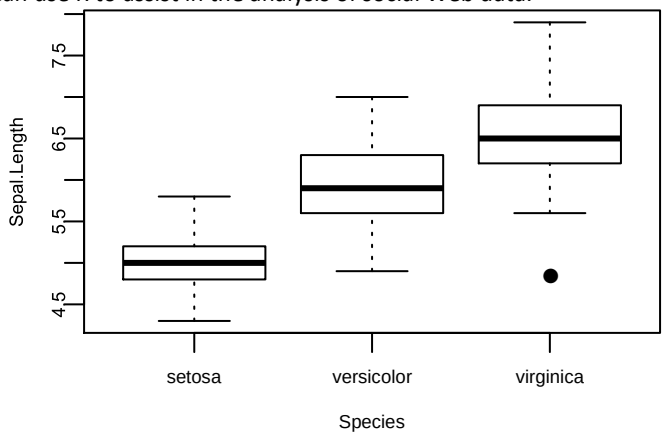

Figure 2. A box plot

Next Week. Introduction to R programming and data structures.