A Math-Fearing Social Scientist's Basic R Toolkit: Scraping, Content and Network Analysis

Pieter E. Stek

Postdoctoral Scholar, Research Center

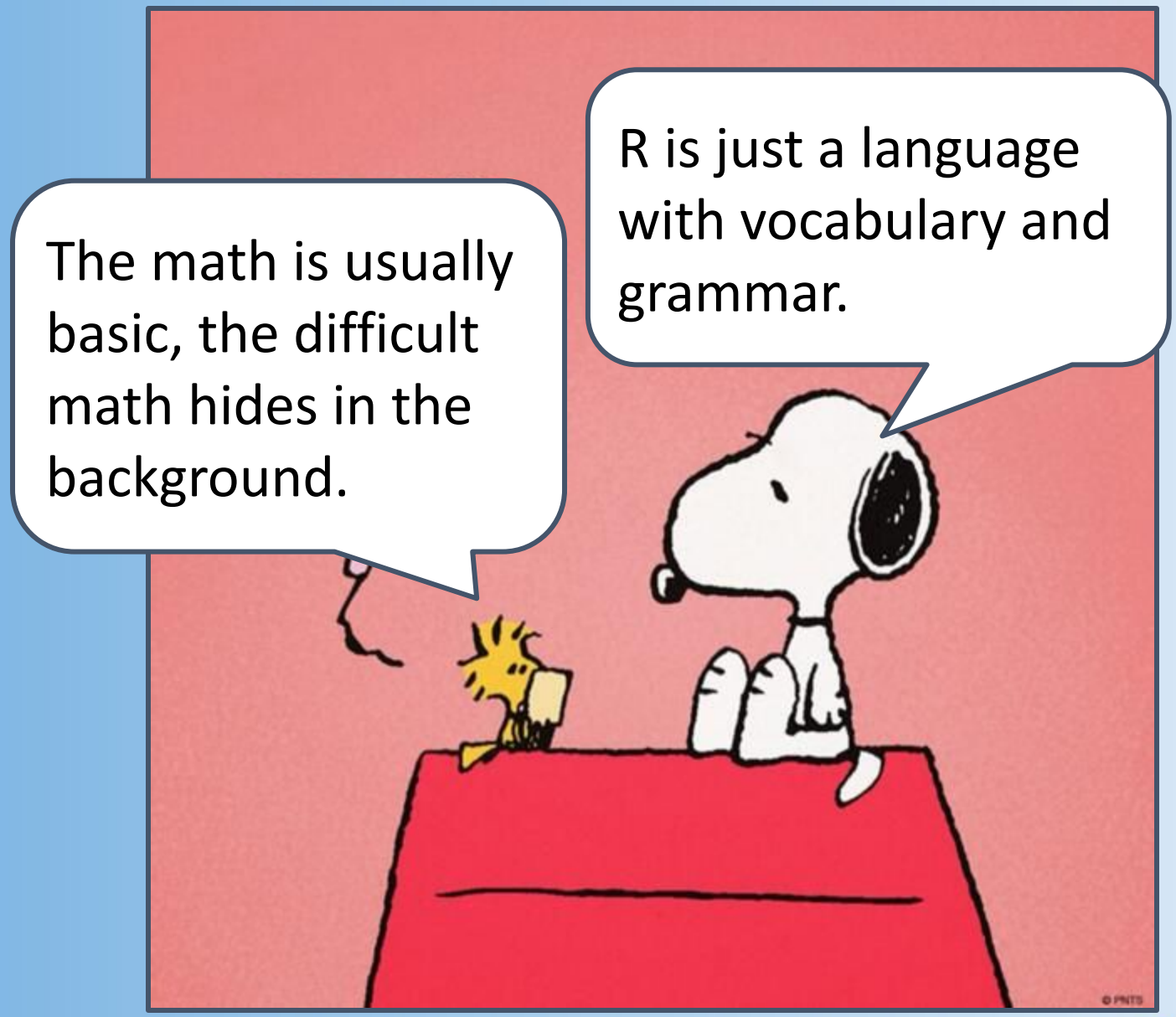
Asia School of Business, Kuala Lumpur, Malaysia

CONFERENCE 2022

https://pstek.nl/2022/r-conference

Slides and R Code

CONFERENCE 2022


R is just a language with vocabulary and grammar.
The math is usually basic, the difficult math hides in the background.

An R Toolbox

Web Scraping
Example 1: news using Google News

Text Analysis
Example 2: news topic analysis

Network Analysis

Example 3: social network from "A Tale of Two Cities"

Tool #1: Web Scraping (Getting Text Data)

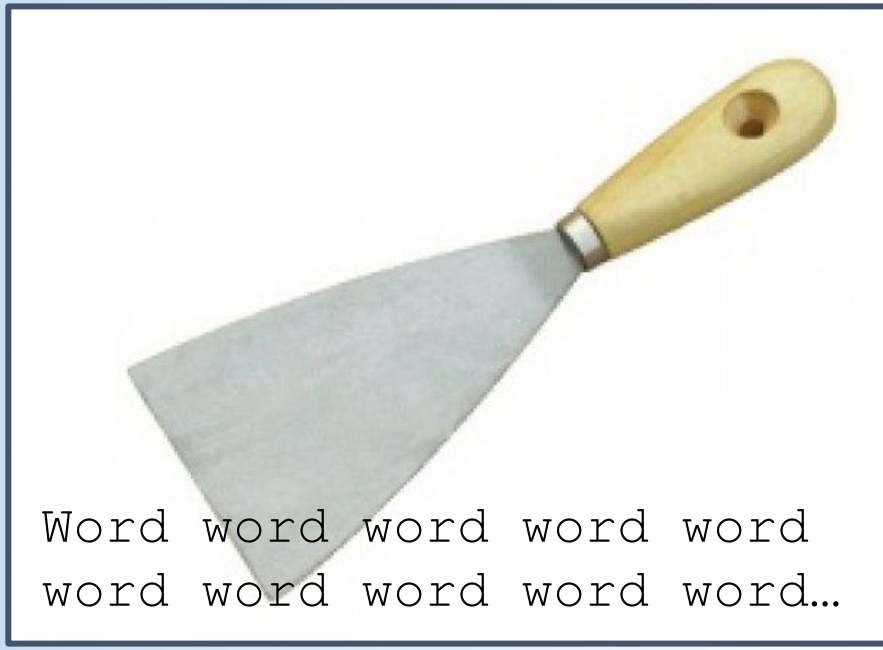
Goal: Scrape news articles using Google News (and later analyze them)

R Packages:

tidyRSS

rvest

Note: Web scraping may violate terms of service and copyright.



Basic Concepts before we Begin

RSS (Really Simple Syndication) is used for news, blogs, podcasts, etc. ("web feeds")

- We use an RSS feed to get a list of news articles.

HTML (HyperText Markup Language) is used to display web pages in browsers.

- We will scrape text from HTML.

What is Google News?

https://news.google.com

Lets you search news through queries.

Example:

```
"Boris Johnson" AND Ukraine site:bbc.co.uk
```

Google News also has RSS feeds!

What can you Scrape?

Anything with digital text.

News: titles, articles, paragraphs

Social media: Twitter, YouTube comments, etc.

Others: Books, scientific articles, patents, etc.

Useful to Know: Loops

A loop is a procedure for a computer program to run through, typically when doing something repetitive… like writing "I love R-Conference 2022" 100 times.

```
for(n in 1:100){
    print('I love R-Conference 2022')
}
```

Tool #2: Text Analysis

Goal: Quantitatively analyze text to understand it better

R Packages:

quanteda

quanteda.textplots

Terminology

Corpus: your text

Cleaning: removing irrelevant text

Tokens: individual words/topics in your text

Frequency analysis: counting words

Co-word analysis: counting if words occur together

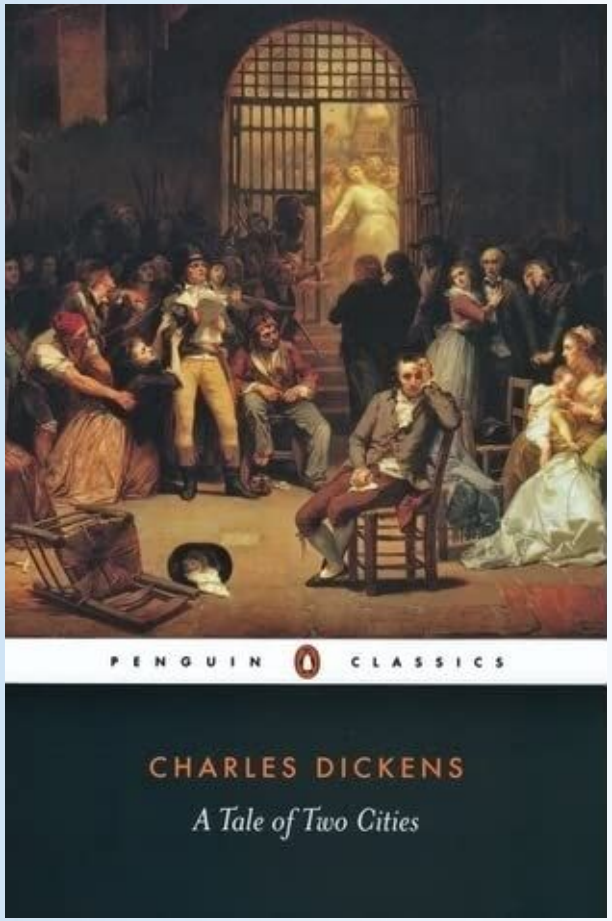
Tool #3: Network Analysis

Goal: Find out who the most important character in a novel is?

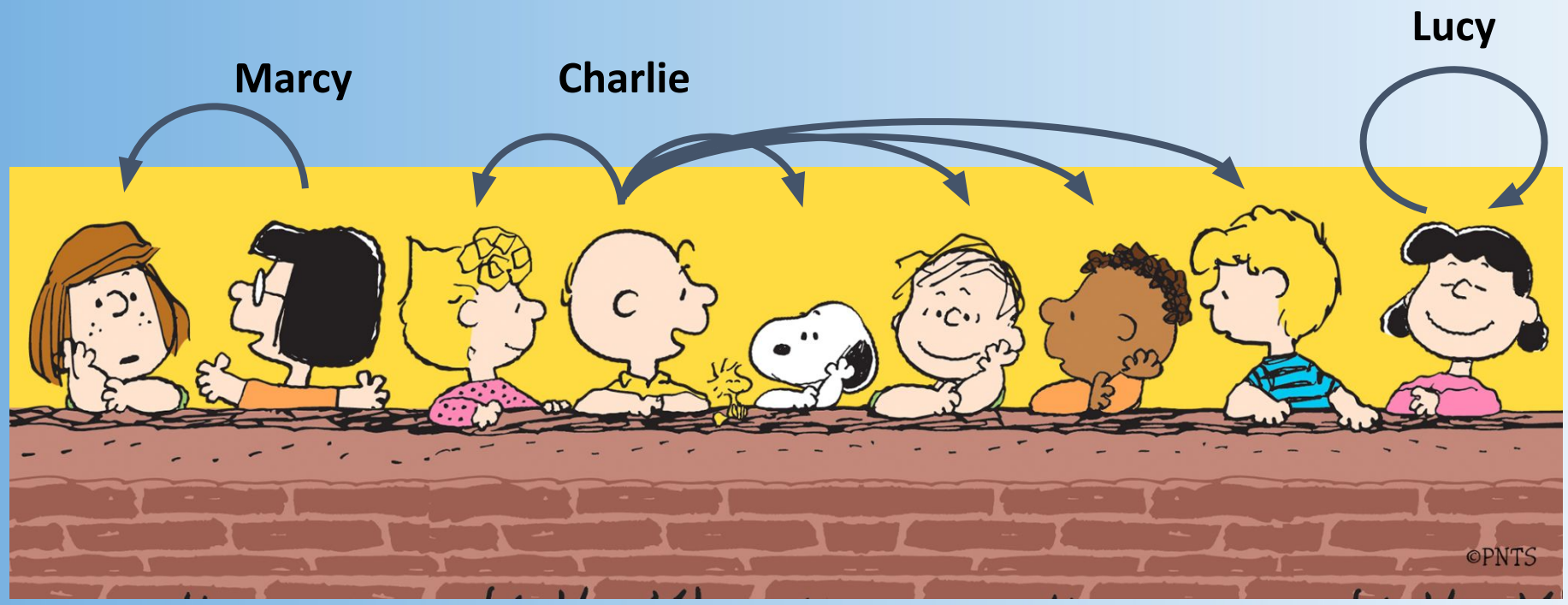
R Packages:

igraph

(quanteda)



Peanuts Communication Network



Terminology

Node: the "thing" in your network

Edge: connection between "things"

Directed/undirected: is there a direction of the edge or not?

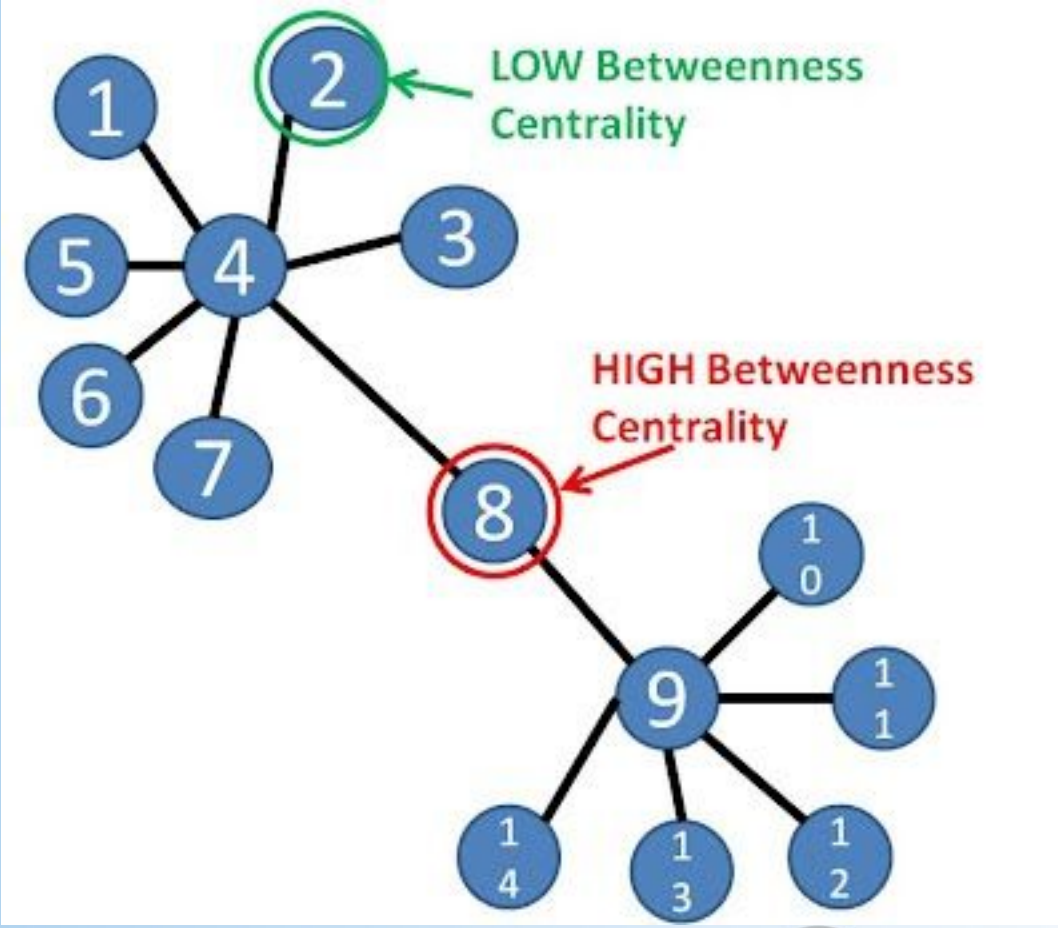
Centrality: measure of how "important" the node is. Includes **degree**, **betweenness**, etc.

Network Analysis Example

(4, 8 & 9) High betweenness centrality **(structural holes)**

(4 & 9) High degree centrality

Image source: https://sites.google.com/site/bsmithactivity3/betweenness-centrality



Thank you!

Slides and code:
https://pstek.nl/2022/r-conference