# Big Data and Machine Learning

Tom Falcone, Mathematics / Computer Science Teacher

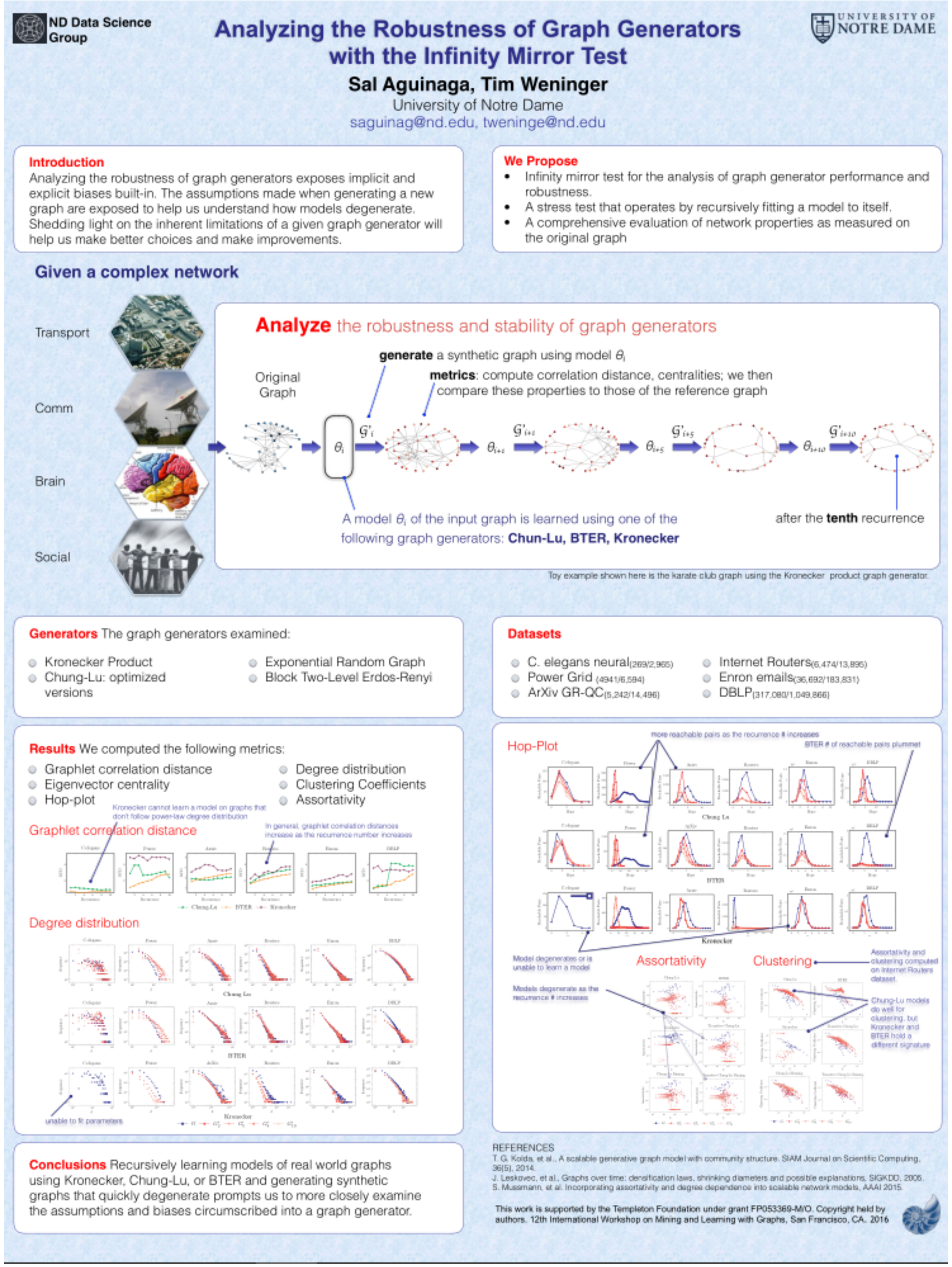
La Lumiere School

Summer 2016 RET Computing

# You need to know
# What I Did This Summer

**iCeNSA:**

**International Center for Network Science and Applications**



ND Data Science Group

UNIVERSITY OF NOTRE DAME

# Analyzing the Robustness of Graph Generators with the Infinity Mirror Test

**Sal Aguinaga, Tim Weninger**
University of Notre Dame
saguinag@nd.edu, tweninge@nd.edu

**Introduction**
Analyzing the robustness of graph generators exposes implicit and explicit biases built-in. The assumptions made when generating a new graph are exposed to help us understand how models degenerate. Shedding light on the inherent limitations of a given graph generator will help us make better choices and make improvements.

**We Propose**
- Infinity mirror test for the analysis of graph generator performance and robustness.
- A stress test that operates by recursively fitting a model to itself.
- A comprehensive evaluation of network properties as measured on the original graph

## Given a complex network

Transport
Comm
Brain
Social

**Analyze** the robustness and stability of graph generators

**generate** a synthetic graph using model $\theta_i$

**metrics**: compute correlation distance, centralities; we then compare these properties to those of the reference graph

A model $\theta_i$ of the input graph is learned using one of the following graph generators: **Chun-Lu, BTER, Kronecker**

after the **tenth** recurrence

Toy example shown here is the karate club graph using the Kronecker product graph generator.

**Generators** The graph generators examined:
- Kronecker Product
- Chung-Lu: optimized versions
- Exponential Random Graph
- Block Two-Level Erdos-Renyi

**Datasets**
- C. elegans neural(269/2,965)
- Power Grid (4941/6,594)
- ArXiv GR-QC(5,242/14,496)
- Internet Routers(6,474/13,895)
- Enron emails(36,692/183,831)
- DBLP(317,080/1,049,866)

**Results** We computed the following metrics:
- Graphlet correlation distance
- Eigenvector centrality
- Hop-plot
- Degree distribution
- Clustering Coefficients
- Assortativity

Kronecker cannot learn a model on graphs that don't follow power-law degree distribution

Graphlet correlation distance

In general, graphlet correlation distances increase as the recurrence number increases

Degree distribution

unable to fit parameters

Hop-Plot

more reachable pairs as the recurrence # increases

BTER # of reachable pairs plummet

Model degenerates or is unable to learn a model

Assortativity

Clustering

Assortativity and clustering computed on Internet Routers dataset.

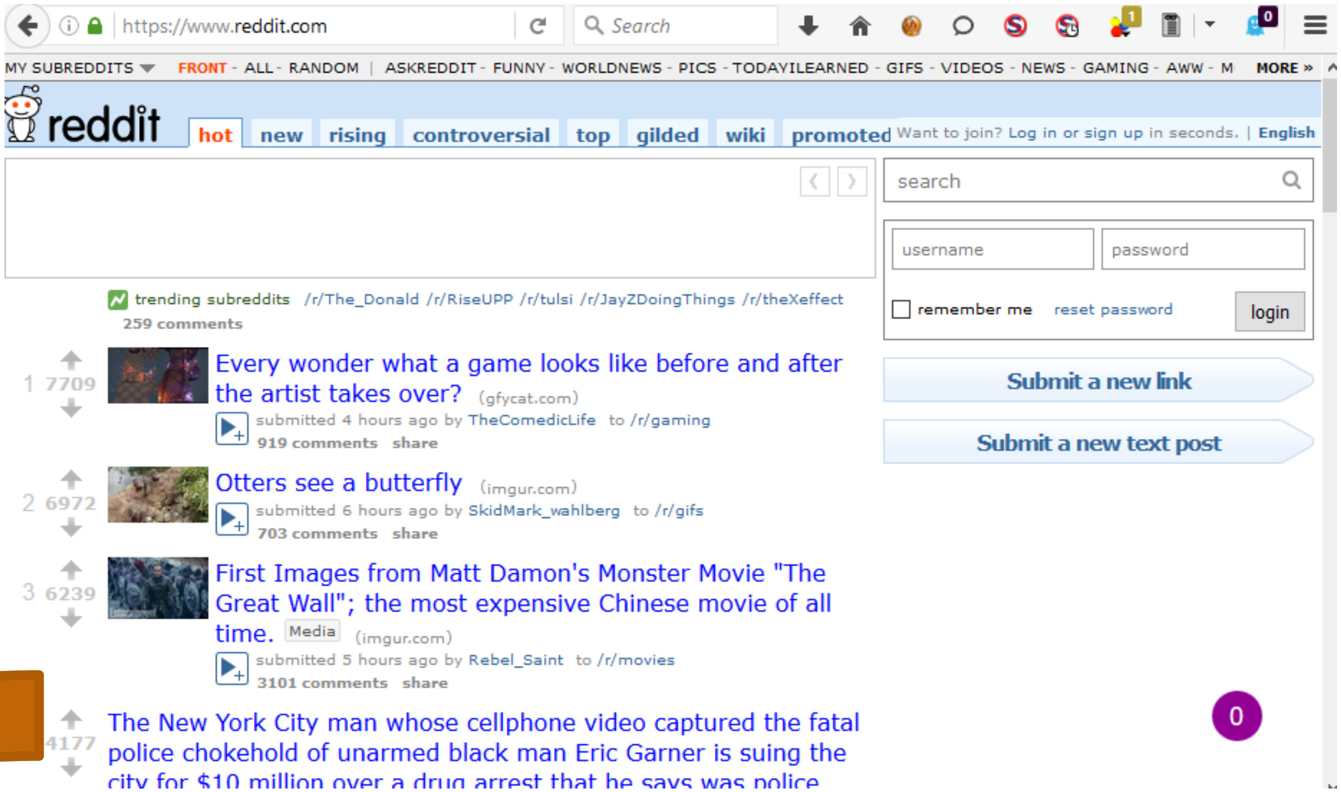
Models degenerate as the recurrence # increases

Chung-Lu models do well for clustering, but Kronecker and BTER hold a different signature

**Conclusions** Recursively learning models of real world graphs using Kronecker, Chung-Lu, or BTER and generating synthetic graphs that quickly degenerate prompts us to more closely examine the assumptions and biases circumscribed into a graph generator.

REFERENCES
T. G. Kolda, et al., A scalable generative graph model with community structure. SIAM Journal on Scientific Computing, 36(5), 2014.
J. Leskovec, et al., Graphs over time: densification laws, shrinking diameters and possible explanations. SIGKDD, 2005.
S. Mussmann, et al. Incorporating assortativity and degree dependence into scalable network models, AAAI 2015.

This work is supported by the Templeton Foundation under grant FP053369-M/O. Copyright held by authors. 12th International Workshop on Mining and Learning with Graphs, San Francisco, CA. 2016



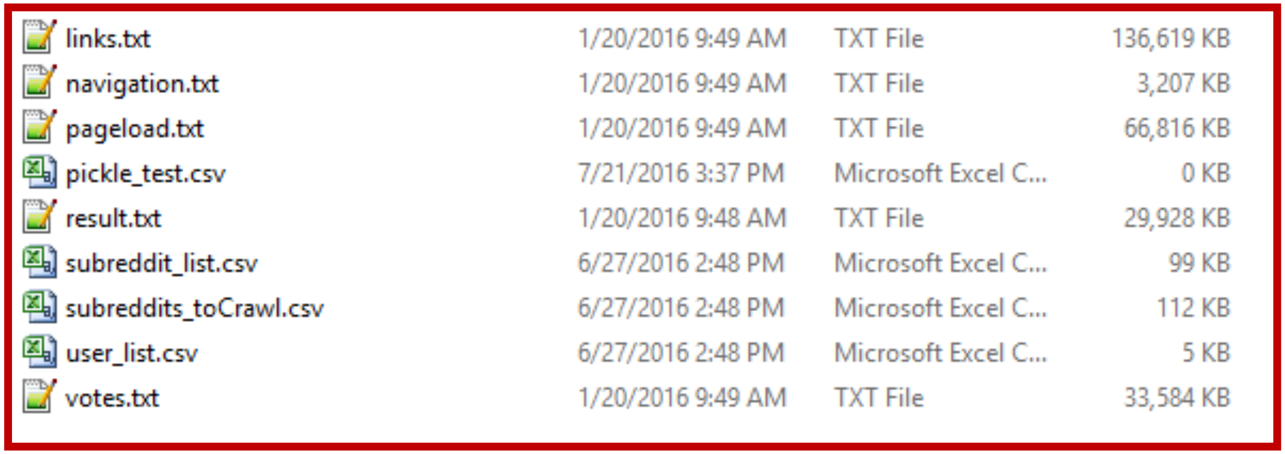
trending subreddits /r/The_Donald /r/RiseUPP /r/tulsi /r/JayZDoingThings /r/theXeffect
259 comments

1 7709 Every wonder what a game looks like before and after the artist takes over? (gfycat.com)
submitted 4 hours ago by TheComedicLife to /r/gaming
919 comments share

2 6972 Otters see a butterfly (imgur.com)
submitted 6 hours ago by SkidMark_wahlberg to /r/gifs
703 comments share

3 6239 First Images from Matt Damon's Monster Movie "The Great Wall"; the most expensive Chinese movie of all time. Media (imgur.com)
submitted 5 hours ago by Rebel_Saint to /r/movies
3101 comments share

4177 The New York City man whose cellphone video captured the fatal police chokehold of unarmed black man Eric Garner is suing the city for $10 million over a drug arrest that he says was police



| | | | |
|---|---|---|---|
| links.txt | 1/20/2016 9:49 AM | TXT File | 136,619 KB |
| navigation.txt | 1/20/2016 9:49 AM | TXT File | 3,207 KB |
| pageload.txt | 1/20/2016 9:49 AM | TXT File | 66,816 KB |
| pickle_test.csv | 7/21/2016 3:37 PM | Microsoft Excel C... | 0 KB |
| result.txt | 1/20/2016 9:48 AM | TXT File | 29,928 KB |
| subreddit_list.csv | 6/27/2016 2:48 PM | Microsoft Excel C... | 99 KB |
| subreddits_toCrawl.csv | 6/27/2016 2:48 PM | Microsoft Excel C... | 112 KB |
| user_list.csv | 6/27/2016 2:48 PM | Microsoft Excel C... | 5 KB |
| votes.txt | 1/20/2016 9:49 AM | TXT File | 33,584 KB |

# Evidence of Online Performance Deterioration in User Sessions on Reddit

Philipp Singer[a,b,*], Emilio Ferrara[c], Farshad Kooti[c], Markus Strohmaier[a,b], and Kristina Lerman[c]

[a]GESIS - Leibniz Institute for the Social Sciences
[b]University of Koblenz
[c]University of Southern California
[*]philipp.singer@gesis.org

## Abstract

This article presents evidence of performance deterioration in online user sessions quantified by studying a massive dataset containing over 55 million comments posted on Reddit in April 2015. After segmenting the sessions (i.e., periods of activity without a prolonged break) depending on their intensity (i.e., how many posts users produced during sessions), we observe a general decrease in the quality of comments produced by users over the course of sessions. We propose mixed-effects models that capture the impact of session intensity on comments, including their length, quality, and the responses they generate from the community. Our findings suggest performance deterioration: Sessions of increasing intensity are associated with the production of shorter, progressively less complex comments, which receive declining quality scores (as rated by other users), and are less and less engaging (i.e., they attract fewer responses). Our contribution evokes a connection between cognitive and attention dynamics and the usage of online social peer production platforms, specifically the effects of deterioration of user performance.

## Introduction

Performance deterioration following a period of sustained mental effort has been documented in settings that include student performance [1], driving [2], data entry [3], and exerting self-control [4]. Although the mechanisms for deteriorating performance are still debated [5, 6, 7], deterioration has been shown to be accompanied by physiological brain changes [8, 9, 10], suggesting a cognitive origin, whether due to mental fatigue, boredom, or strategic choices to limit attention. Outside of vigilance tasks, however, relatively little is known about whether and how this phenomenon affects online behavior. As our society

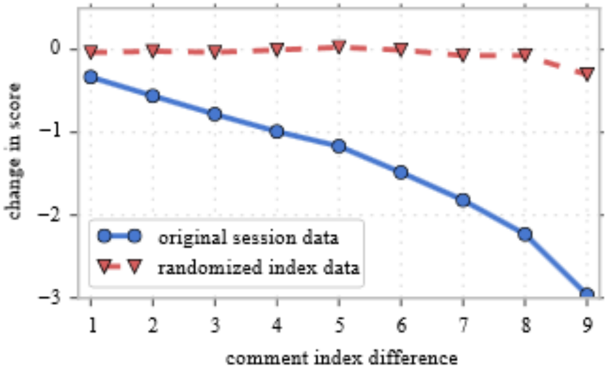
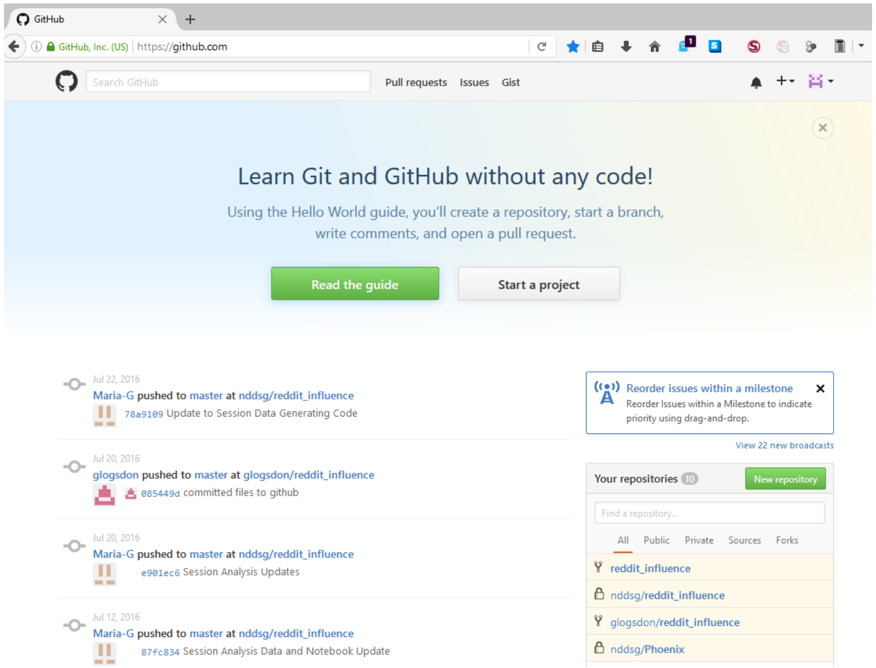


Figure 1: *Performance of comments within sessions.* We show the average Reddit score for comments in sessions of length 10 (original session data, blue solid line). The average rating of each comment decreases starkly, by about 0.3 points for each comment after the first one in the session. This suggests the presence of (super linear) performance deterioration throughout user sessions. The effect disappears in randomized data having shuffled comments within sessions (red dashed line).



Learn Git and GitHub without any code!

Using the Hello World guide, you'll create a repository, start a branch, write comments, and open a pull request.

Jul 22, 2016
Maria-G pushed to master at nddsg/reddit_influence
78a9109 Update to Session Data Generating Code

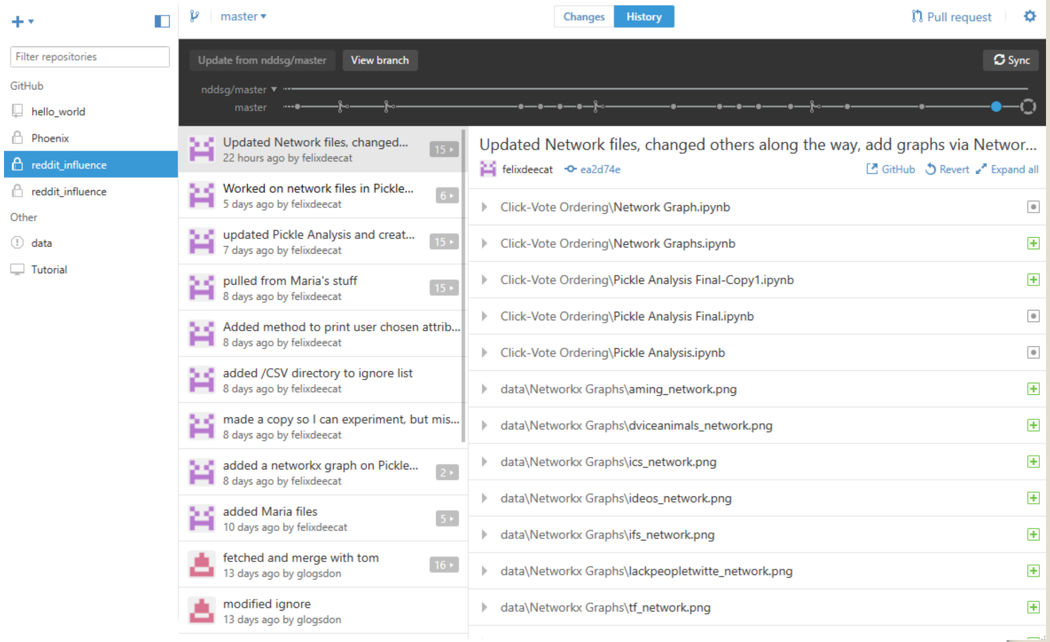
Jul 20, 2016
glogsdon pushed to master at glogsdon/reddit_influence
085449d committed files to github

Jul 20, 2016
Maria-G pushed to master at nddsg/reddit_influence
e901ec6 Session Analysis Updates

Jul 12, 2016
Maria-G pushed to master at nddsg/reddit_influence
87fc834 Session Analysis Data and Notebook Update

Reorder issues within a milestone
Reorder Issues within a Milestone to indicate priority using drag-and-drop.
View 22 new broadcasts

Your repositories 10
- reddit_influence
- nddsg/reddit_influence
- glogsdon/reddit_influence
- nddsg/Phoenix



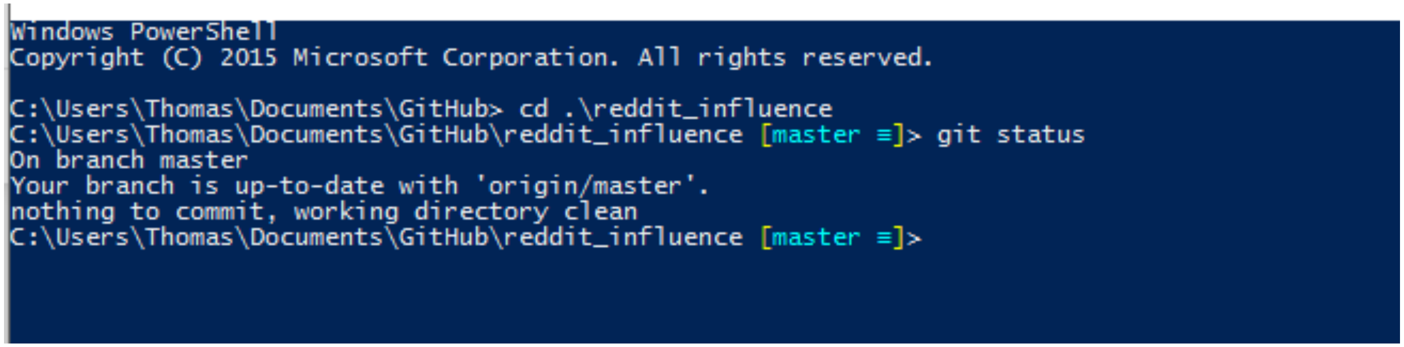
GitHub
- hello_world
- Phoenix
- reddit_influence
- reddit_influence

Other
- data
- Tutorial

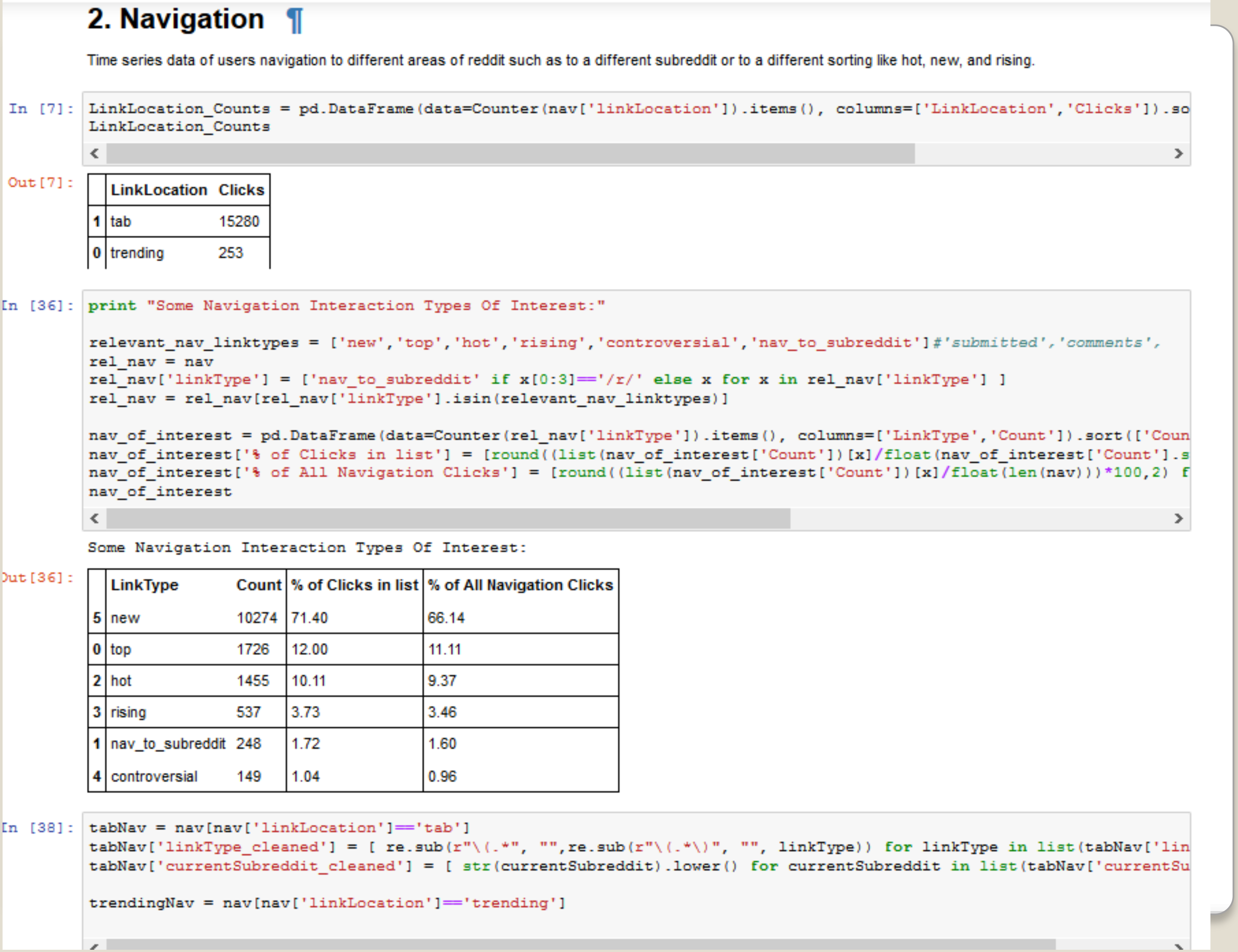
Updated Network files, changed... 22 hours ago by felixdeecat 15
Worked on network files in Pickle... 5 days ago by felixdeecat 6
updated Pickle Analysis and creat... 7 days ago by felixdeecat 15
pulled from Maria's stuff 8 days ago by felixdeecat 15
Added method to print user chosen attrib... 8 days ago by felixdeecat
added /CSV directory to ignore list 8 days ago by felixdeecat
made a copy so I can experiment, but mis... 8 days ago by felixdeecat
added a networkx graph on Pickle... 8 days ago by felixdeecat 2
added Maria files 10 days ago by felixdeecat 5
fetched and merge with tom 13 days ago by glogsdon 16
modified ignore 13 days ago by glogsdon

Updated Network files, changed others along the way, add graphs via Networ...
felixdeecat ea2d74e

- Click-Vote Ordering\Network Graph.ipynb
- Click-Vote Ordering\Network Graphs.ipynb
- Click-Vote Ordering\Pickle Analysis Final-Copy1.ipynb
- Click-Vote Ordering\Pickle Analysis Final.ipynb
- Click-Vote Ordering\Pickle Analysis.ipynb
- data\Networkx Graphs\aming_network.png
- data\Networkx Graphs\dviceanimals_network.png
- data\Networkx Graphs\ics_network.png
- data\Networkx Graphs\ideos_network.png
- data\Networkx Graphs\ifs_network.png
- data\Networkx Graphs\lackpeopletwitte_network.png
- data\Networkx Graphs\tf_network.png



```
Windows PowerShell
Copyright (C) 2015 Microsoft Corporation. All rights reserved.

C:\Users\Thomas\Documents\GitHub> cd .\reddit_influence
C:\Users\Thomas\Documents\GitHub\reddit_influence [master ≡]> git status
On branch master
Your branch is up-to-date with 'origin/master'.
nothing to commit, working directory clean
C:\Users\Thomas\Documents\GitHub\reddit_influence [master ≡]>
```



## 2. Navigation ¶

Time series data of users navigation to different areas of reddit such as to a different subreddit or to a different sorting like hot, new, and rising.

In [7]:

```
LinkLocation_Counts = pd.DataFrame(data=Counter(nav['linkLocation']).items(), columns=['LinkLocation','Clicks']).so
LinkLocation_Counts
```

Out[7]:

|   | LinkLocation | Clicks |
|---|---|---|
| **1** | tab | 15280 |
| **0** | trending | 253 |

In [36]:

```
print "Some Navigation Interaction Types Of Interest:"

relevant_nav_linktypes = ['new','top','hot','rising','controversial','nav_to_subreddit']#'submitted','comments',
rel_nav = nav
rel_nav['linkType'] = ['nav_to_subreddit' if x[0:3]=='/r/' else x for x in rel_nav['linkType'] ]
rel_nav = rel_nav[rel_nav['linkType'].isin(relevant_nav_linktypes)]

nav_of_interest = pd.DataFrame(data=Counter(rel_nav['linkType']).items(), columns=['LinkType','Count']).sort(['Coun
nav_of_interest['% of Clicks in list'] = [round((list(nav_of_interest['Count'])[x]/float(nav_of_interest['Count'].s
nav_of_interest['% of All Navigation Clicks'] = [round((list(nav_of_interest['Count'])[x]/float(len(nav)))*100,2) f
nav_of_interest
```

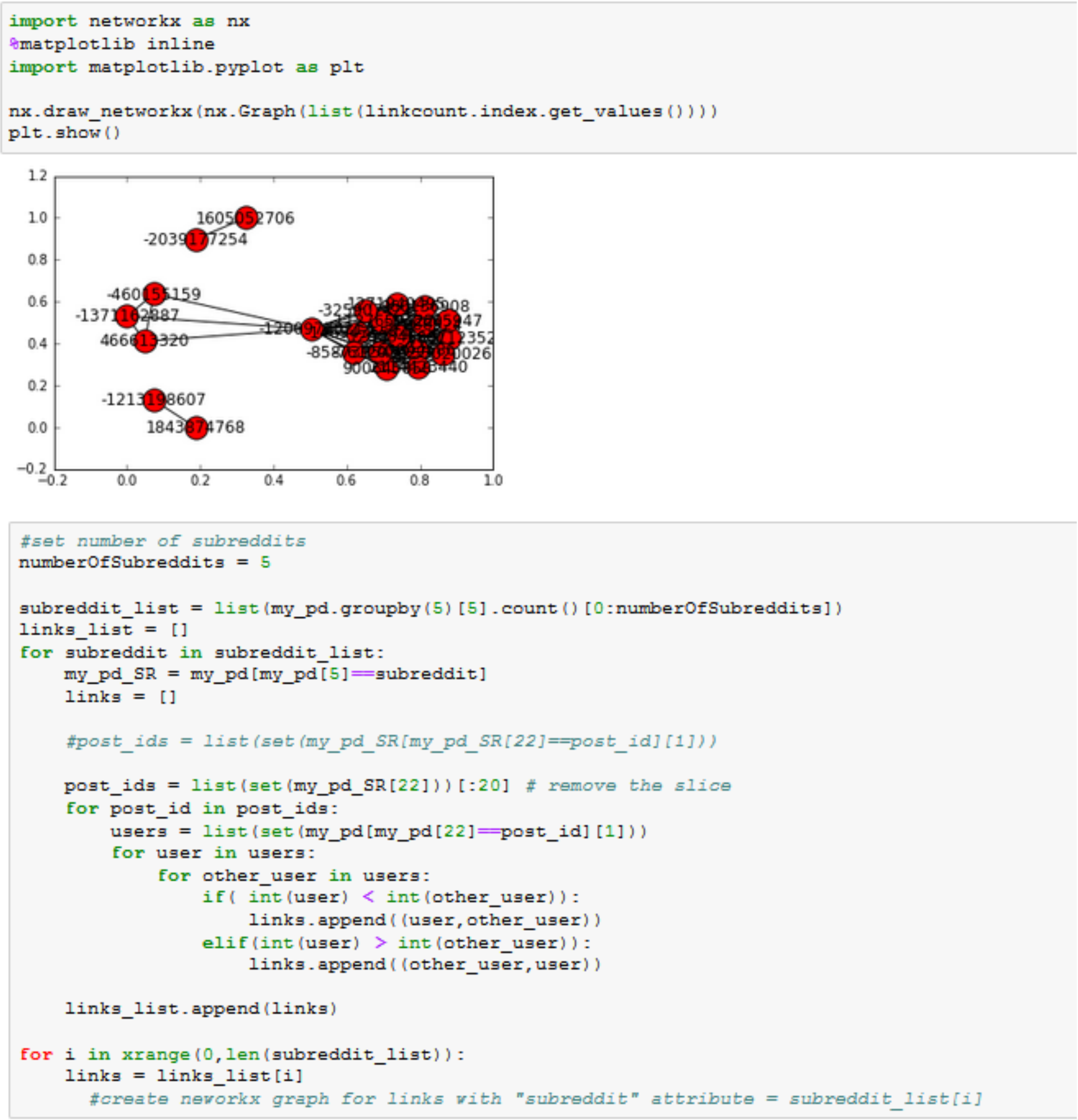
```
Some Navigation Interaction Types Of Interest:
```

Out[36]:

|   | LinkType | Count | % of Clicks in list | % of All Navigation Clicks |
|---|---|---|---|---|
| **5** | new | 10274 | 71.40 | 66.14 |
| **0** | top | 1726 | 12.00 | 11.11 |
| **2** | hot | 1455 | 10.11 | 9.37 |
| **3** | rising | 537 | 3.73 | 3.46 |
| **1** | nav_to_subreddit | 248 | 1.72 | 1.60 |
| **4** | controversial | 149 | 1.04 | 0.96 |

In [38]:

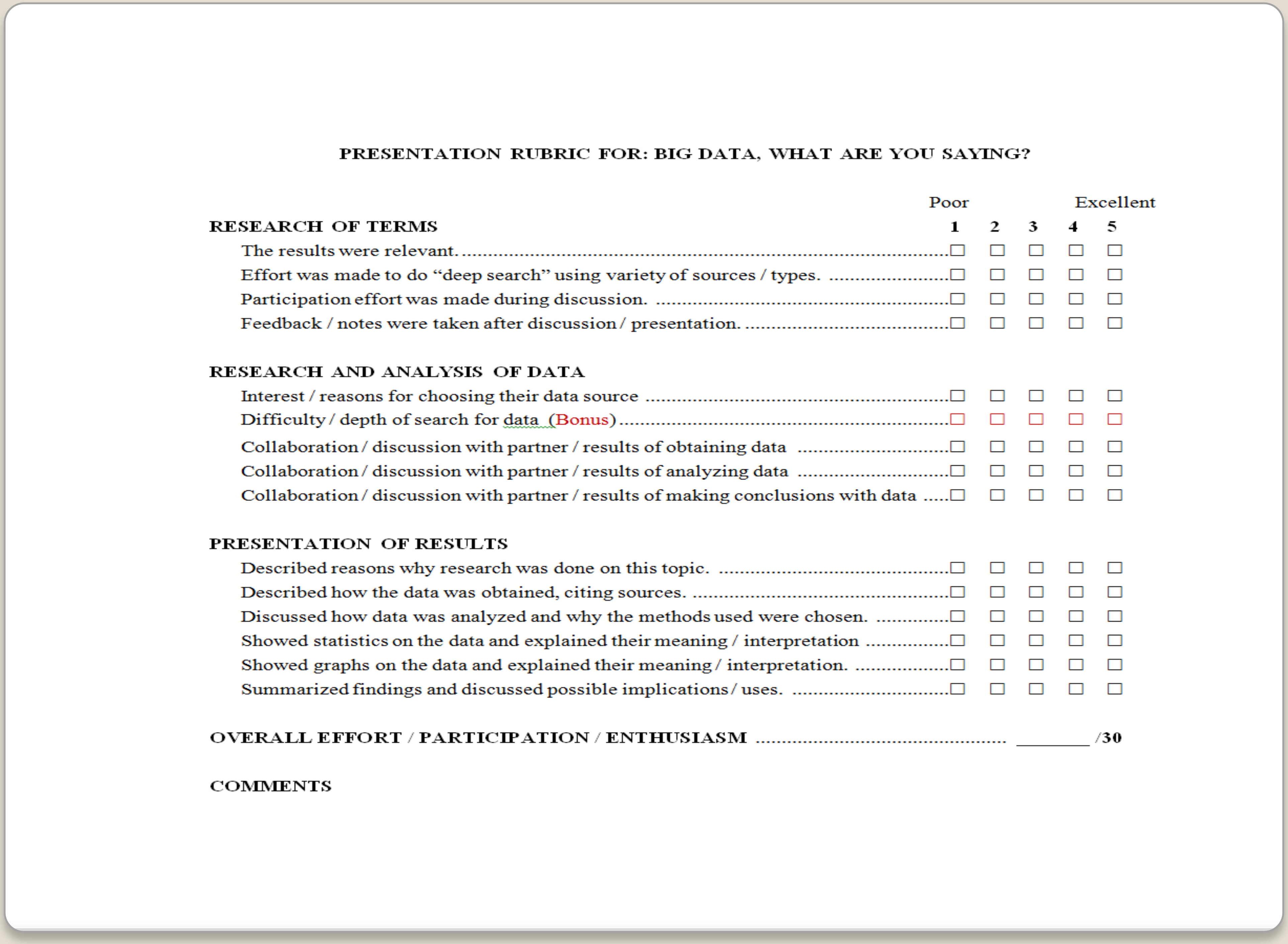
```
tabNav = nav[nav['linkLocation']=='tab']
tabNav['linkType_cleaned'] = [ re.sub(r"\(.*", "",re.sub(r"\(.*\)", "", linkType)) for linkType in list(tabNav['lin
tabNav['currentSubreddit_cleaned'] = [ str(currentSubreddit).lower() for currentSubreddit in list(tabNav['currentSu

trendingNav = nav[nav['linkLocation']=='trending']
```



```
import networkx as nx
%matplotlib inline
import matplotlib.pyplot as plt

nx.draw_networkx(nx.Graph(list(linkcount.index.get_values())))
plt.show()
```

```
#set number of subreddits
numberOfSubreddits = 5

subreddit_list = list(my_pd.groupby(5)[5].count()[0:numberOfSubreddits])
links_list = []
for subreddit in subreddit_list:
    my_pd_SR = my_pd[my_pd[5]==subreddit]
    links = []

    #post_ids = list(set(my_pd_SR[my_pd_SR[22]==post_id][1]))

    post_ids = list(set(my_pd_SR[22]))[:20] # remove the slice
    for post_id in post_ids:
        users = list(set(my_pd[my_pd[22]==post_id][1]))
        for user in users:
            for other_user in users:
                if( int(user) < int(other_user)):
                    links.append((user,other_user))
                elif(int(user) > int(other_user)):
                    links.append((other_user,user))

    links_list.append(links)

for i in xrange(0,len(subreddit_list)):
    links = links_list[i]
      #create neworkx graph for links with "subreddit" attribute = subreddit_list[i]
```

# TRANSLATING OUR STUDIES TO THE CLASSROOM



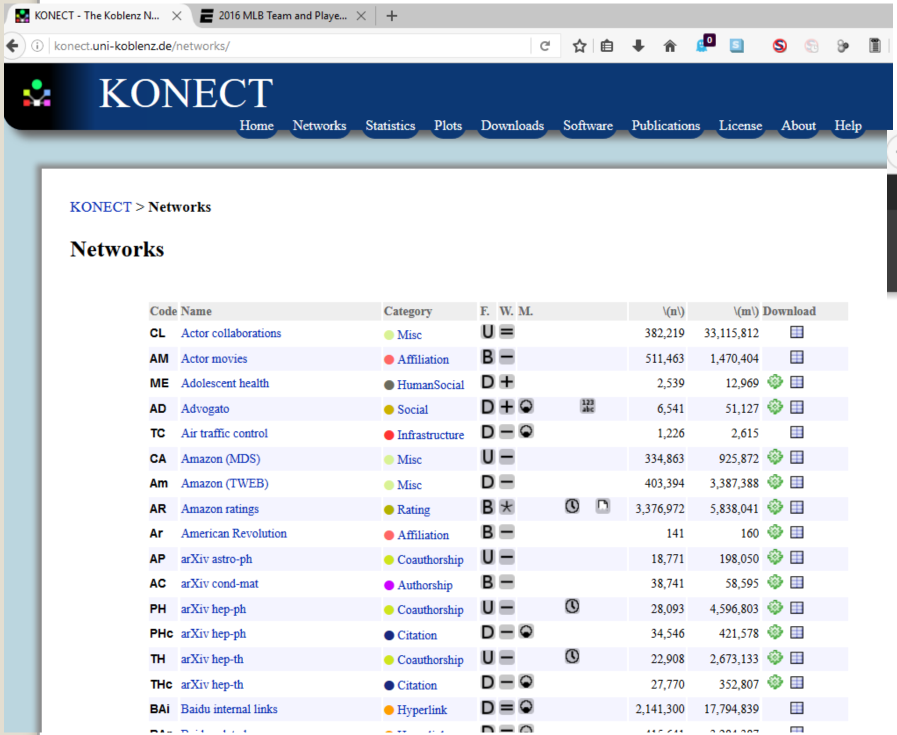
## PRESENTATION RUBRIC FOR: BIG DATA, WHAT ARE YOU SAYING?

| | Poor | | | | Excellent |
|---|---|---|---|---|---|
| **RESEARCH OF TERMS** | **1** | **2** | **3** | **4** | **5** |
| The results were relevant. | ☐ | ☐ | ☐ | ☐ | ☐ |
| Effort was made to do "deep search" using variety of sources / types. | ☐ | ☐ | ☐ | ☐ | ☐ |
| Participation effort was made during discussion. | ☐ | ☐ | ☐ | ☐ | ☐ |
| Feedback / notes were taken after discussion / presentation. | ☐ | ☐ | ☐ | ☐ | ☐ |
| **RESEARCH AND ANALYSIS OF DATA** | | | | | |
| Interest / reasons for choosing their data source | ☐ | ☐ | ☐ | ☐ | ☐ |
| Difficulty / depth of search for data (Bonus) | ☐ | ☐ | ☐ | ☐ | ☐ |
| Collaboration / discussion with partner / results of obtaining data | ☐ | ☐ | ☐ | ☐ | ☐ |
| Collaboration / discussion with partner / results of analyzing data | ☐ | ☐ | ☐ | ☐ | ☐ |
| Collaboration / discussion with partner / results of making conclusions with data | ☐ | ☐ | ☐ | ☐ | ☐ |
| **PRESENTATION OF RESULTS** | | | | | |
| Described reasons why research was done on this topic. | ☐ | ☐ | ☐ | ☐ | ☐ |
| Described how the data was obtained, citing sources. | ☐ | ☐ | ☐ | ☐ | ☐ |
| Discussed how data was analyzed and why the methods used were chosen. | ☐ | ☐ | ☐ | ☐ | ☐ |
| Showed statistics on the data and explained their meaning / interpretation | ☐ | ☐ | ☐ | ☐ | ☐ |
| Showed graphs on the data and explained their meaning / interpretation. | ☐ | ☐ | ☐ | ☐ | ☐ |
| Summarized findings and discussed possible implications / uses. | ☐ | ☐ | ☐ | ☐ | ☐ |

**OVERALL EFFORT / PARTICIPATION / ENTHUSIASM** ............................................ ________ /30

**COMMENTS**

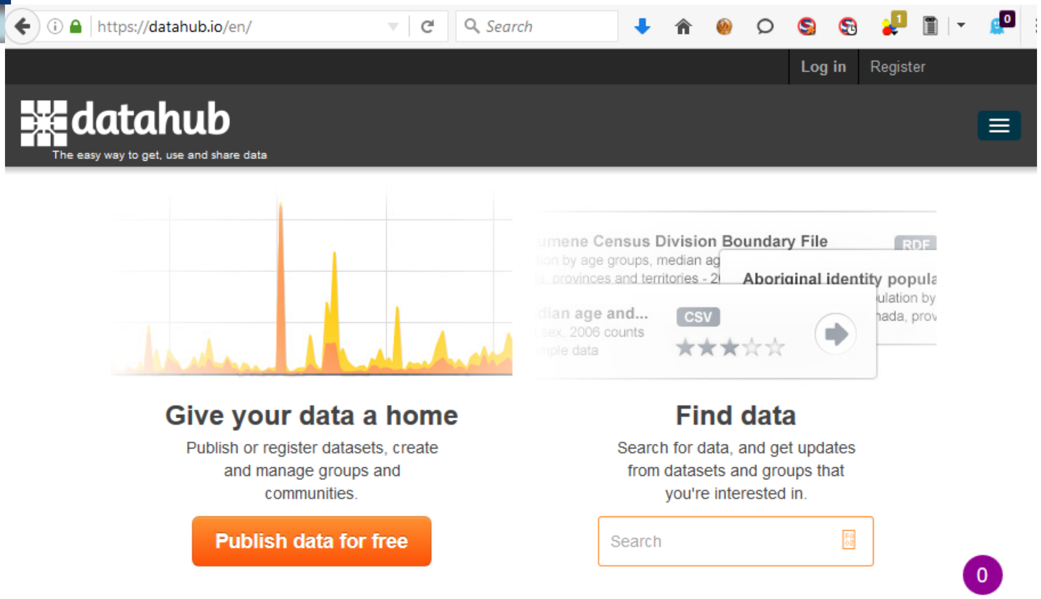
# Data Sources


KONECT - The Koblenz N...
2016 MLB Team and Playe...
konect.uni-koblenz.de/networks/
KONECT
Home
Networks
Statistics
Plots
Downloads
Software
Publications
License
About
Help

KONECT > Networks

## Networks

| Code | Name | Category | F. | W. | M. | \(n\) | \(m\) | Download |
|---|---|---|---|---|---|---|---|---|
| CL | Actor collaborations | Misc | U | = | | 382,219 | 33,115,812 | ▦ |
| AM | Actor movies | Affiliation | B | − | | 511,463 | 1,470,404 | ▦ |
| ME | Adolescent health | HumanSocial | D | + | | 2,539 | 12,969 | ▦ |
| AD | Advogato | Social | D | + | | 6,541 | 51,127 | ▦ |
| TC | Air traffic control | Infrastructure | D | − | | 1,226 | 2,615 | ▦ |
| CA | Amazon (MDS) | Misc | U | − | | 334,863 | 925,872 | ▦ |
| Am | Amazon (TWEB) | Misc | D | − | | 403,394 | 3,387,388 | ▦ |
| AR | Amazon ratings | Rating | B | * | | 3,376,972 | 5,838,041 | ▦ |
| Ar | American Revolution | Affiliation | B | − | | 141 | 160 | ▦ |
| AP | arXiv astro-ph | Coauthorship | U | − | | 18,771 | 198,050 | ▦ |
| AC | arXiv cond-mat | Authorship | B | − | | 38,741 | 58,595 | ▦ |
| PH | arXiv hep-ph | Coauthorship | U | − | | 28,093 | 4,596,803 | ▦ |
| PHc | arXiv hep-ph | Citation | D | − | | 34,546 | 421,578 | ▦ |
| TH | arXiv hep-th | Coauthorship | U | − | | 22,908 | 2,673,133 | ▦ |
| THc | arXiv hep-th | Citation | D | − | | 27,770 | 352,807 | ▦ |
| BAi | Baidu internal links | Hyperlink | D | = | | 2,141,300 | 17,794,839 | ▦ |

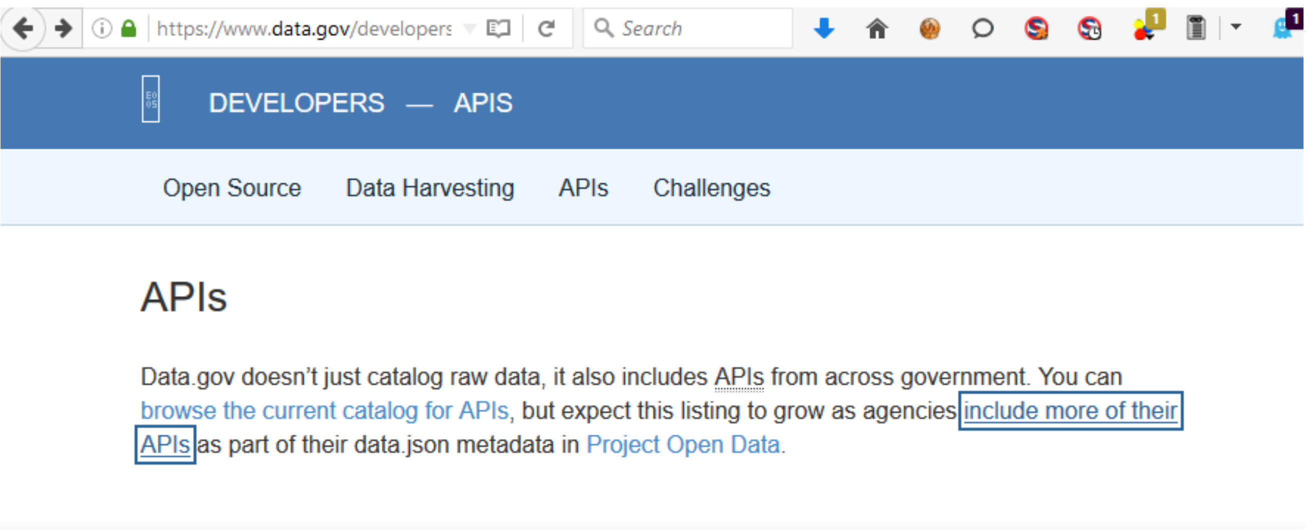


### Give your data a home

Publish or register datasets, create and manage groups and communities.

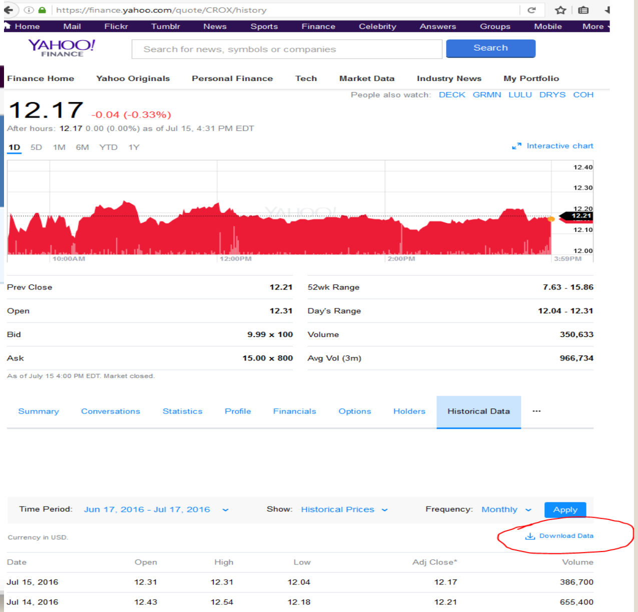
### Find data

Search for data, and get updates from datasets and groups that you're interested in.



## APIs

Data.gov doesn't just catalog raw data, it also includes APIs from across government. You can browse the current catalog for APIs, but expect this listing to grow as agencies include more of their APIs as part of their data.json metadata in Project Open Data.



12.17 -0.04 (-0.33%)

After hours: 12.17 0.00 (0.00%) as of Jul 15, 4:31 PM EDT

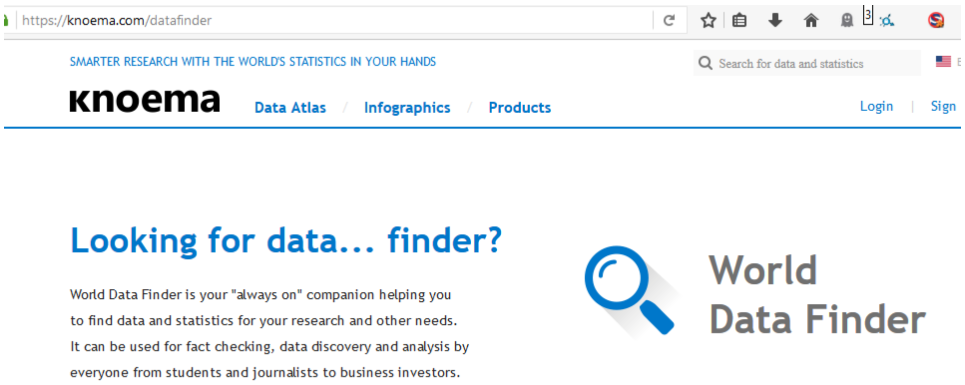
| | | | |
|---|---|---|---|
| Prev Close | 12.21 | 52wk Range | 7.63 - 15.86 |
| Open | 12.31 | Day's Range | 12.04 - 12.31 |
| Bid | 9.99 x 100 | Volume | 350,633 |
| Ask | 15.00 x 800 | Avg Vol (3m) | 966,734 |

As of July 15 4:00 PM EDT. Market closed.

Time Period: Jun 17, 2016 - Jul 17, 2016 Show: Historical Prices Frequency: Monthly Apply

Currency in USD.

| Date | Open | High | Low | Adj Close* | Volume |
|---|---|---|---|---|---|
| Jul 15, 2016 | 12.31 | 12.31 | 12.04 | 12.17 | 386,700 |
| Jul 14, 2016 | 12.43 | 12.54 | 12.18 | 12.21 | 655,400 |


https://knoema.com/datafinder

SMARTER RESEARCH WITH THE WORLD'S STATISTICS IN YOUR HANDS

Search for data and statistics

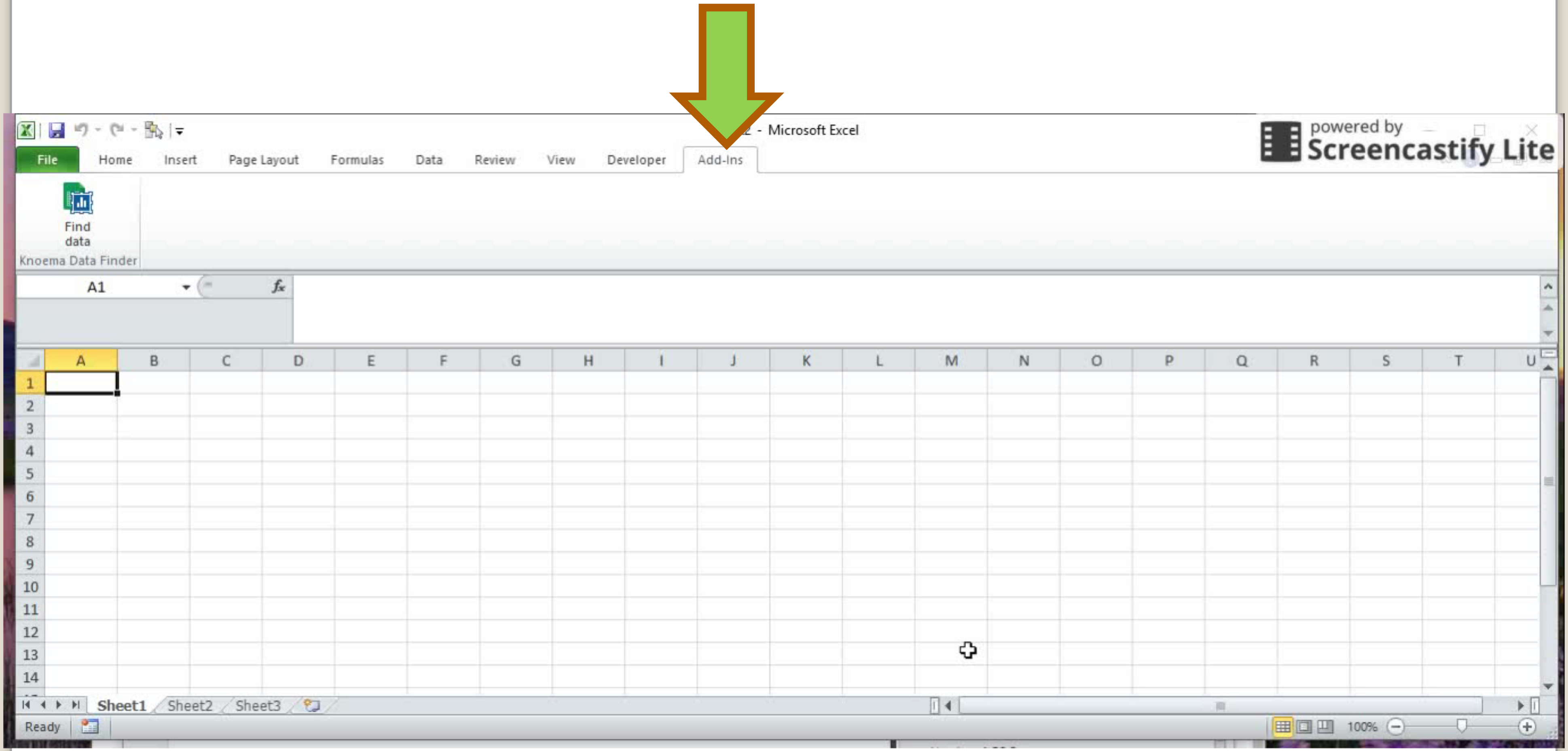
**knoema** Data Atlas / Infographics / Products

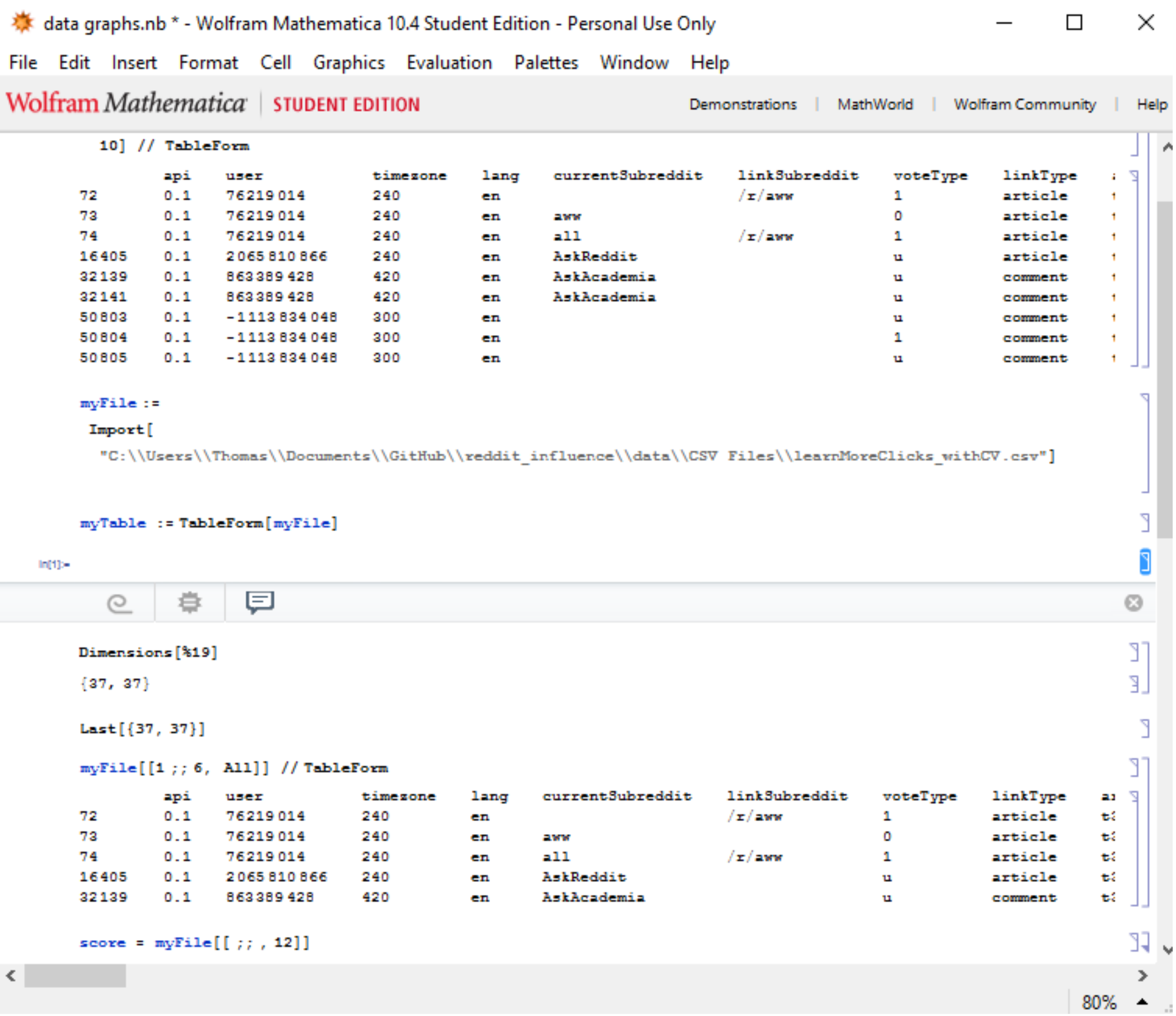
Login | Sign

## Looking for data... finder?

World Data Finder is your "always on" companion helping you to find data and statistics for your research and other needs. It can be used for fact checking, data discovery and analysis by everyone from students and journalists to business investors.

World Data Finder





```
  10] // TableForm
         api   user          timezone   lang   currentSubreddit   linkSubreddit   voteType   linkType
72       0.1   76219014      240        en                        /r/aww          1          article
73       0.1   76219014      240        en     aww                                0          article
74       0.1   76219014      240        en     all                /r/aww          1          article
16405    0.1   2065810866    240        en     AskReddit                          u          article
32139    0.1   863389428     420        en     AskAcademia                        u          comment
32141    0.1   863389428     420        en     AskAcademia                        u          comment
50803    0.1   -1113834048   300        en                                        u          comment
50804    0.1   -1113834048   300        en                                        1          comment
50805    0.1   -1113834048   300        en                                        u          comment

myFile :=
 Import[
  "C:\\Users\\Thomas\\Documents\\GitHub\\reddit_influence\\data\\CSV Files\\learnMoreClicks_withCV.csv"]


myTable := TableForm[myFile]

In[1]:=
```

```
Dimensions[%19]
{37, 37}

Last[{37, 37}]

myFile[[1 ;; 6, All]] // TableForm
         api   user          timezone   lang   currentSubreddit   linkSubreddit   voteType   linkType
72       0.1   76219014      240        en                        /r/aww          1          article
73       0.1   76219014      240        en     aww                                0          article
74       0.1   76219014      240        en     all                /r/aww          1          article
16405    0.1   2065810866    240        en     AskReddit                          u          article
32139    0.1   863389428     420        en     AskAcademia                        u          comment

score = myFile[[ ;; , 12]]
```

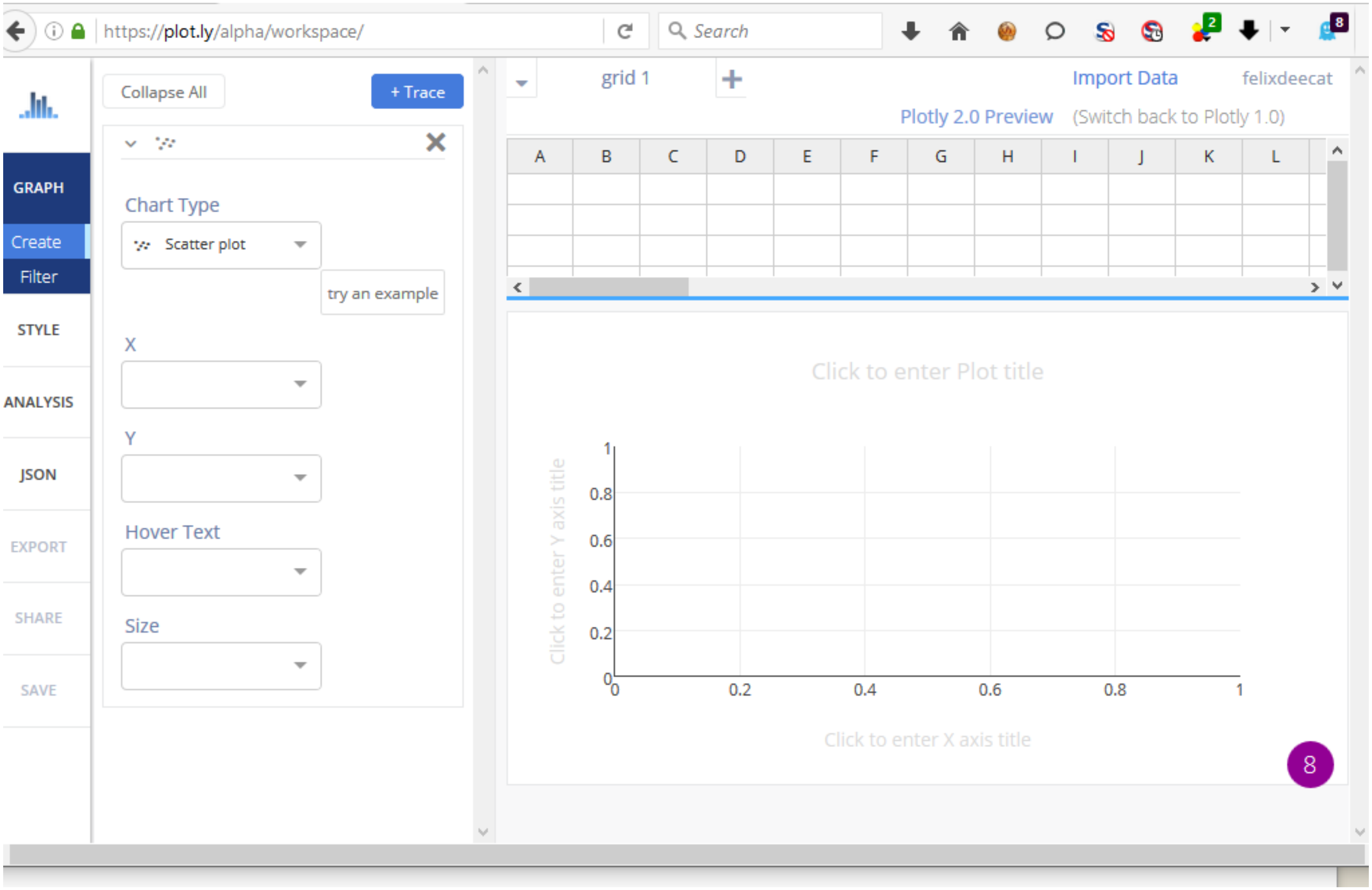
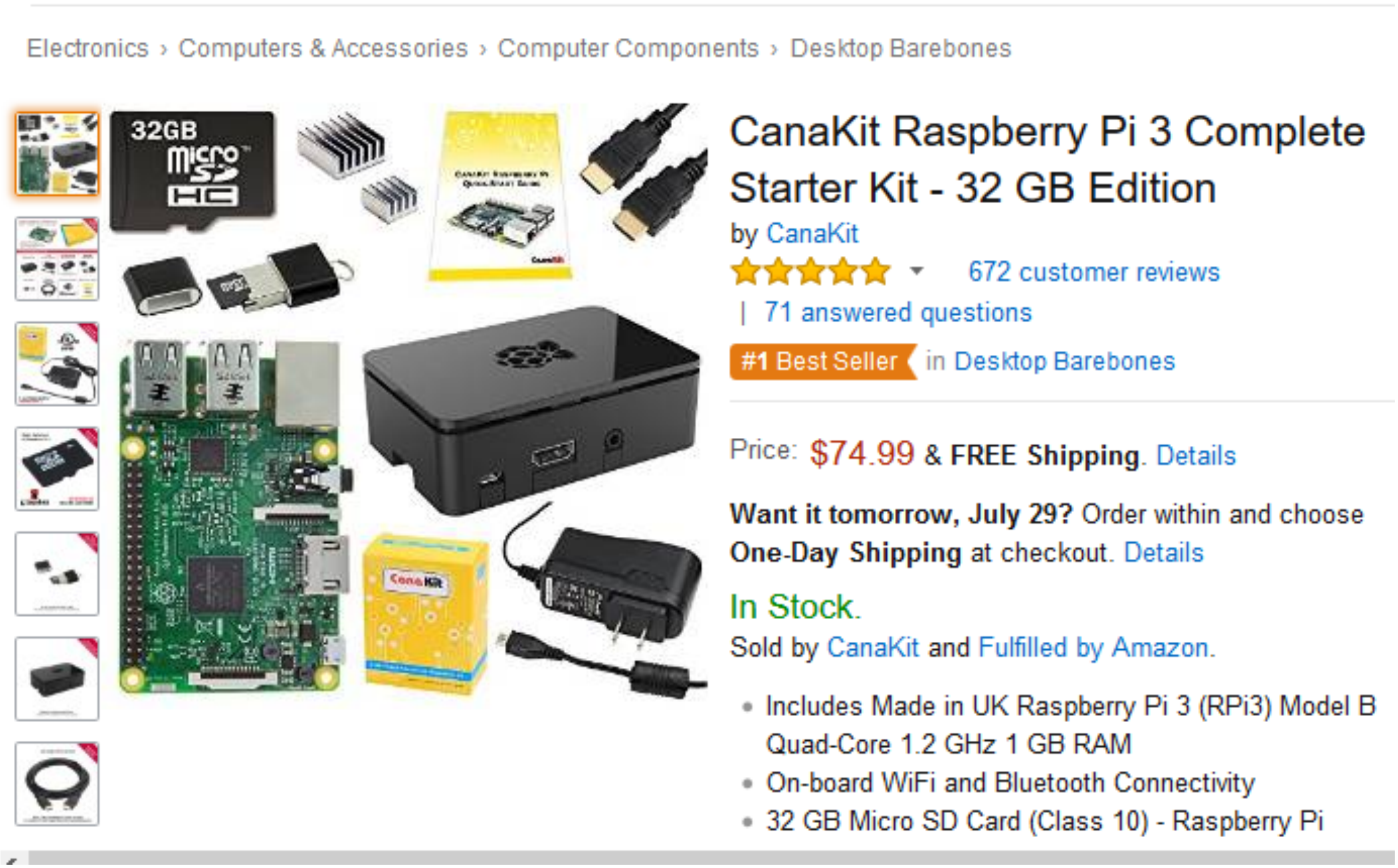


Chart Type
Scatter plot



Electronics › Computers & Accessories › Computer Components › Desktop Barebones

CanaKit Raspberry Pi 3 Complete Starter Kit - 32 GB Edition

by CanaKit

★★★★★ 672 customer reviews

| 71 answered questions

**#1 Best Seller** in Desktop Barebones

Price: $74.99 & **FREE Shipping**. Details

**Want it tomorrow, July 29?** Order within and choose **One-Day Shipping** at checkout. Details

In Stock.

Sold by CanaKit and Fulfilled by Amazon.

- Includes Made in UK Raspberry Pi 3 (RPi3) Model B Quad-Core 1.2 GHz 1 GB RAM
- On-board WiFi and Bluetooth Connectivity
- 32 GB Micro SD Card (Class 10) - Raspberry Pi

# THANK YOU

Michael Niemier
Tim Weninger
Sal Aguinaga
Corey Pennycuff
Maria Glenski
RET Computing Program and its funders and supporters