

# **GE2234 Social Networks**

for Media, Business and Technological Applications

## **Lecture Note 5: Network Data Collection**

By Dr. Wang Xiaohui, Vincent

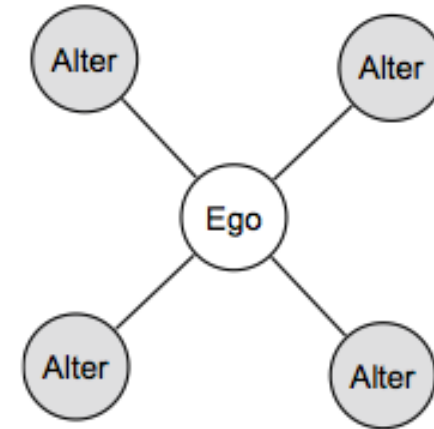
# Recap

- Indegree and Out Degree Centrality [popular]
- Closeness Centrality [graphically centered]
- Betweenness Centrality [gatekeeper]
- Eigenvector Centrality

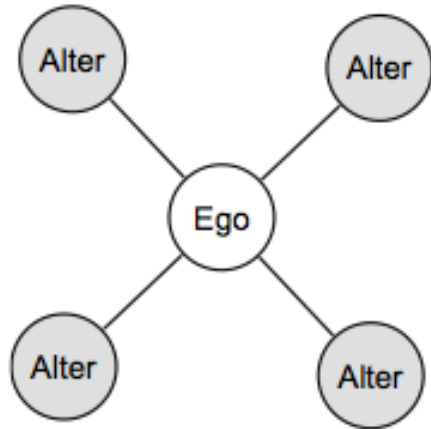
Egocentric, Partial, and Full Networks

# Ego Network

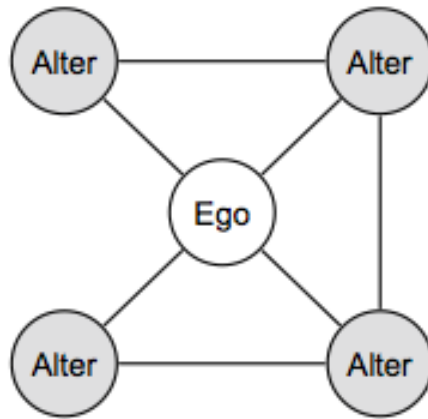
- Ego network
  - **Ego**: the individual that is the focus of attention
  - **Alters**: the people he/she is connected to
  - Example: a network of your personal Facebook friends
- Ego's network is a source of: Information, Social support, Access to resources, etc. All of which can influence Ego's behavior
- Usage: combines the perspective of network analysis with the data of mainstream social science



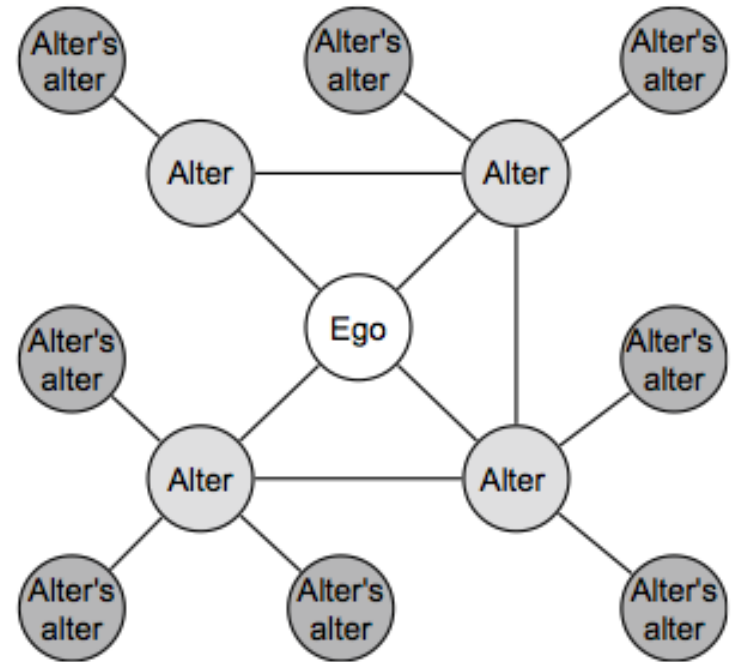
# The depth of ego network



1-Degree Ego Network



1.5-Degree Ego Network

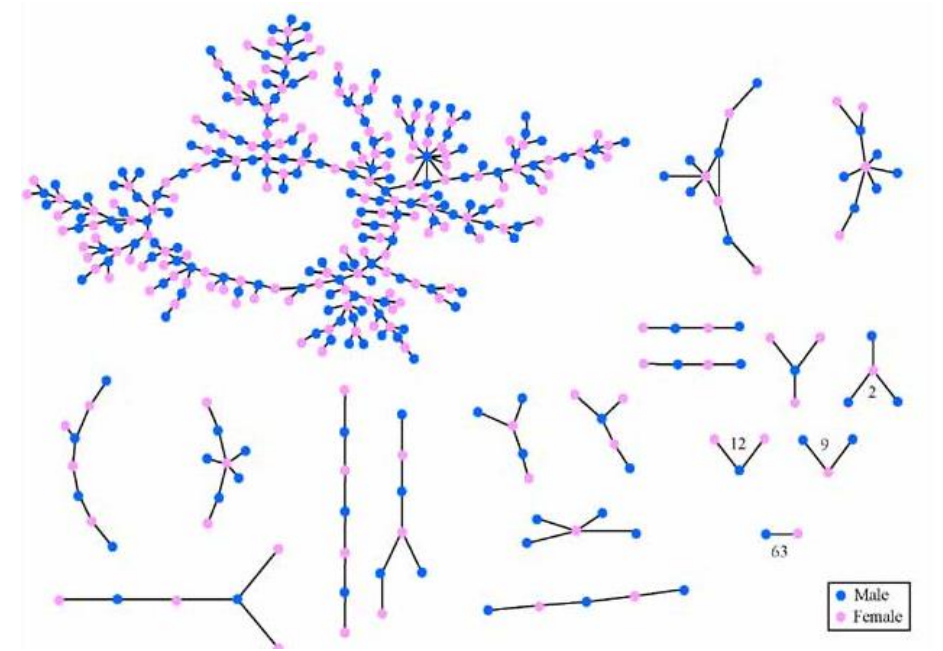


2-Degree Ego Network

# Full and Partial Networks

Full Network: contains all the people or entities of interest and the connections among them.

- Data of all actors within a particular (relevant) boundary
- Never exactly complete (due to missing data), but boundaries are set
  - Co-authorship data among all writers in the social sciences
  - friendships among all students in a classroom

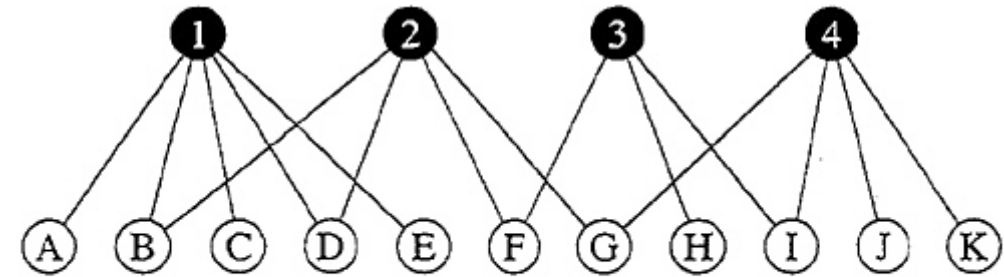


Partial Network: a sample or slice of the full network.

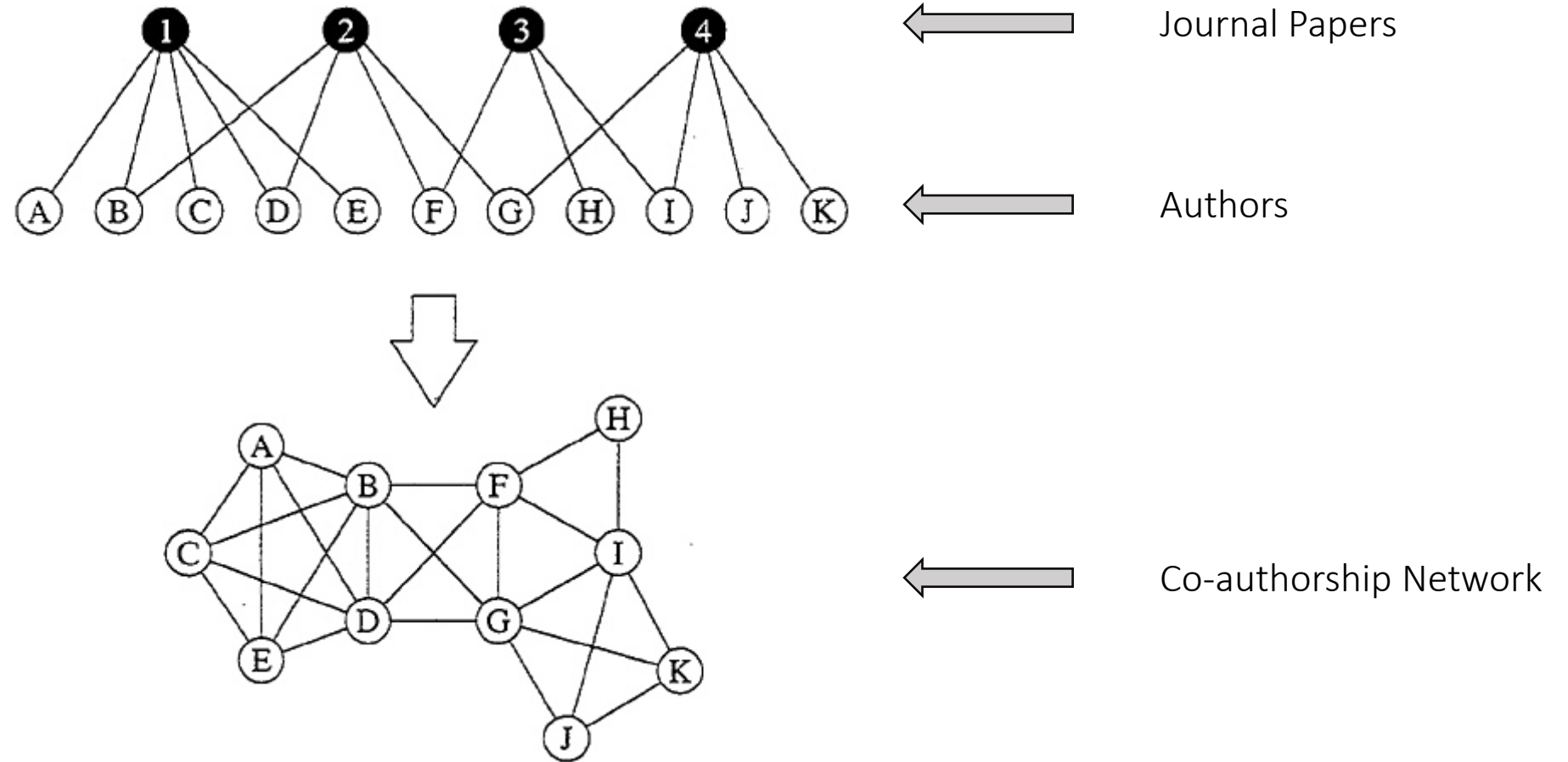
- **Topic centric** (connections between users interested in a topic)
- **Time sliced** (connections existing during a certain period of time)
- **Subgrouping by activities** (e.g., sections of a conference, joining a community)

# Unimodal, Affiliation, and Multimodal Networks

- **Unimodal networks:** only one type of nodes included (e.g., users-users, webpages-webpages)
- **Multimodal networks:** different types of nodes included in the network
  - A common type of multimodal network is Bimodal network (Affiliation network)
- **Bimodal Network**
- Bimodal data represents nodes from two separate classes, where all ties are across classes. Examples:
  - *People* as members of *groups*
  - *People* as authors on *papers*
  - *Words* used often by *people*
  - *Events* in the life history of *people*
- The two modes of the data represent a duality
  - You can project the data as people connected to people through joint membership in a group, or groups connected to groups through common membership.

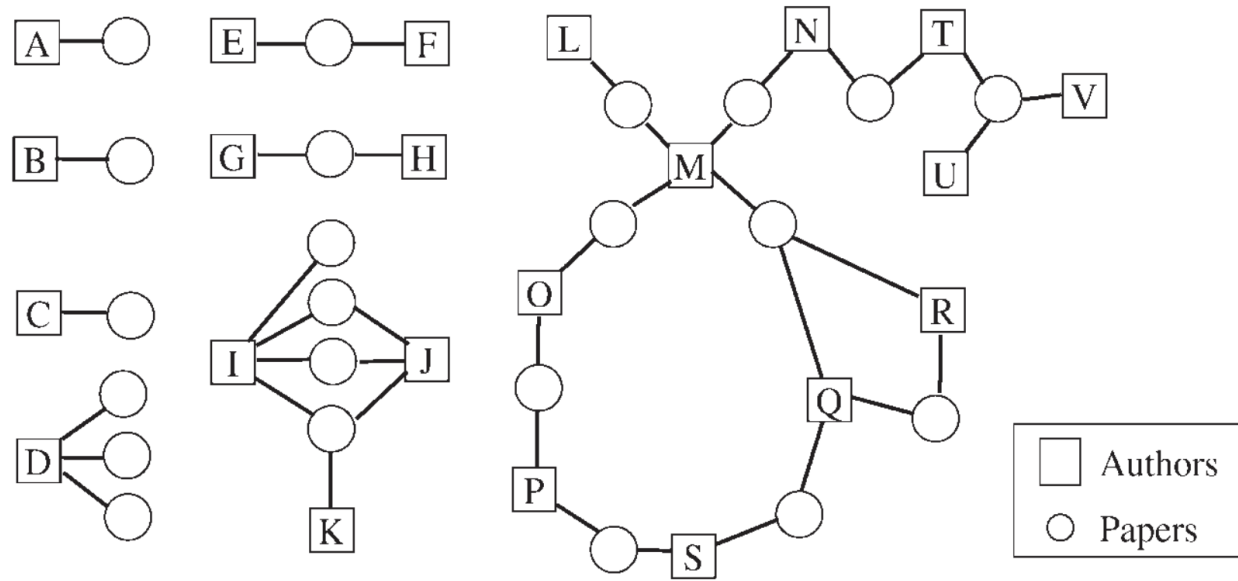


# An Illustration on Multimode to Unimode Transformation

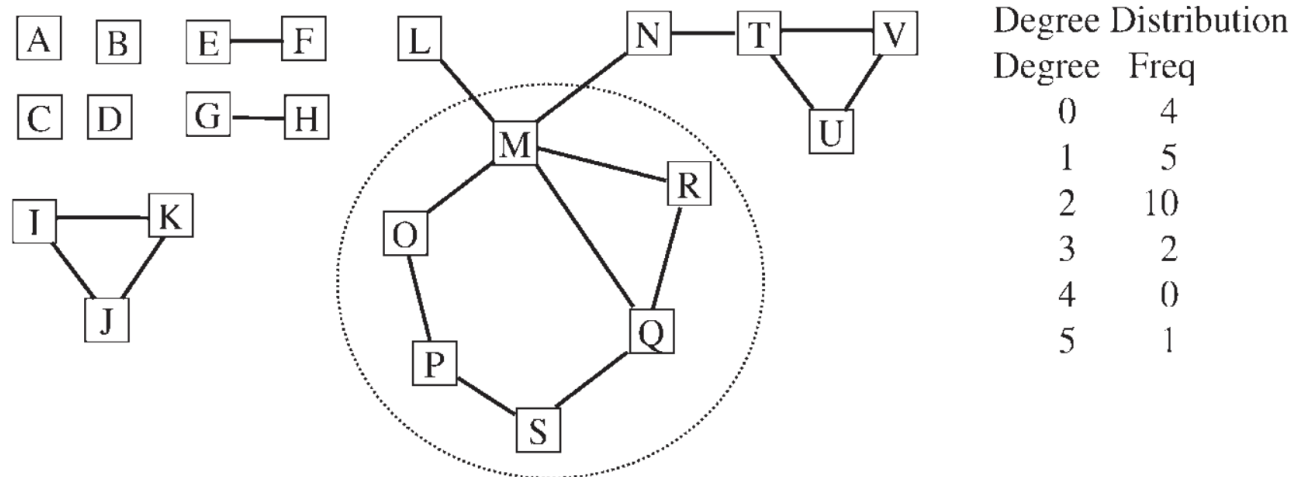




### a) Individual Publications



### b) Collaboration Network



# Data Collection

- Archival Data
- Second-hand data
- Write an FOI Request
- First-hand data – API
- First-hand data – Crawling
- First-hand data – survey

Going Straight to the Sources

# Archival Data

- Records kept in online library because of legal requirements
- Some of the most common source
  - Public records from governmental agencies
    - [Population By-census](#)
  - Research organizations
    - [香港民意研究計劃](#)
  - Health and human service organizations
    - [Health Information National Trends Survey | HINTS](#)
  - Business and industry
- Examples: Chinese Medicine Formulae Images Database




SCM | Library | Main Page |  
All records | Advanced search

Keyword search  "182" record(s) found

Quick search

Astringent formulas | Blood-regulating formulas | Dampclearing formulas  
Digestive formulas | Dryness-moistening formulas | Emetic formulas  
Exterior-releasing formulas | Harmonizing formulas | Heat-clearing formulas  
Phlegm-expelling formulas | Purgative formulas | Qi-regulating formulas  
Resuscitative formulas | Sedative and tranquilizing formulas  
Summer-heat clearing formulas | Tonic formulas | Vermifuge formulas  
Warming interior formulas | Wind-calming formulas

◀◀ ◀ 1 2 3 4 5 6 7 ... ▶ ▶▶

| Numbering | Name                                     | Combination   | Action   | Indication   | Thumbnail   |
|-----------|--|---|--|--|---|
| 1         | Effective Integration Decoction          | Glehniae Radix;<br>Ophiopogonis Radix;<br>Angelicae Sinensis Radix;<br>Rehmanniae Radix; Lycii Fructus; Toosendan Fructus                 | Enriches yin and soothes the liver.                              | Yi Guan Jian is appropriate for patterns of liver-kidney yin deficiency with liver qi constraint characterized by pain in the chest, abdomen, and hypochondriac regions. |    |
| 2         | Nine Immortals Powder                    | Ginseng Radix et Rhizoma; Farfarae Flos; Mori Cortex; Platycodonis Radix; Schisandrae Chinensis Fructus; Asini Corii Colla; Mume Fructus; | Astringes the lung, relieves cough, boosts qi and nourishes yin. | Jiu Xian San is designed for a pattern of chronic cough caused by lung deficiency. The main symptoms include a chronic, unremitting cough that may be accompanied by     |   |
| 3         | Nine Ingredients Notopterygium Decoction | Notopterygii Rhizoma et Radix; Saposhnikovia Radix; Atractylodis Rhizoma; Asari Radix et Rhizoma; Chuanxiong Rhizoma; Angelicae           | Induces sweating, expels dampness, and clears internal heat.     | Jiu Wei Qiang Huo Tang is indicated for an exterior pattern with externally contracted wind-cold-dampness complicated by interior heat. The symptoms                     |  |

Chinese Medicine Formulae Images Database

<https://sys02.lib.hkbu.edu.hk/cmfid/index.asp?lang=eng>

## Official data portals

- [Data.gov.hk](https://data.gov.hk)

## World Bank Open Data

- focus on people's well-being
- can be pulled by country, topic or indicator,
- <https://data.worldbank.org/>

## Our World in Data

- a scholarly database
- focus on global overview of changes in living conditions
- <https://ourworldindata.org/>

## ICPSR: Find & Analyze Data

- research data in social and behavioral sciences
- <https://www.icpsr.umich.edu/icpsrweb/ICPSR/>

## Statista

- business and industry statistical data
- <https://www-statista-com.lib-ezproxy.hkbu.edu.hk/>

# Archival Data

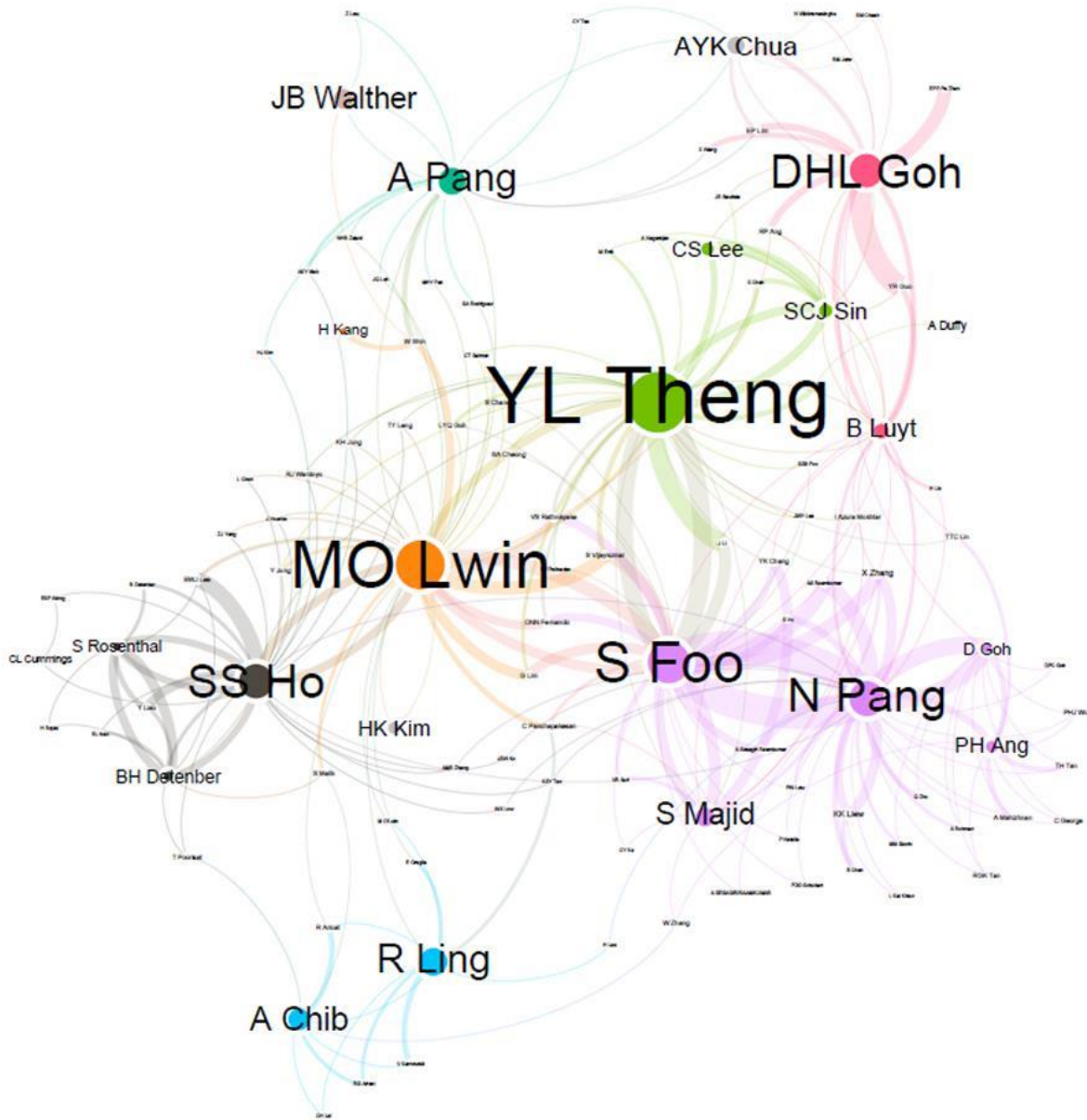
- Some archival sources are ready-to-use
- Some archival sources are indirectly inferred (inferred from raw data)

## Exploring interdependencies in global resource trade



China's resources imports: <https://resourcetrade.earth/>



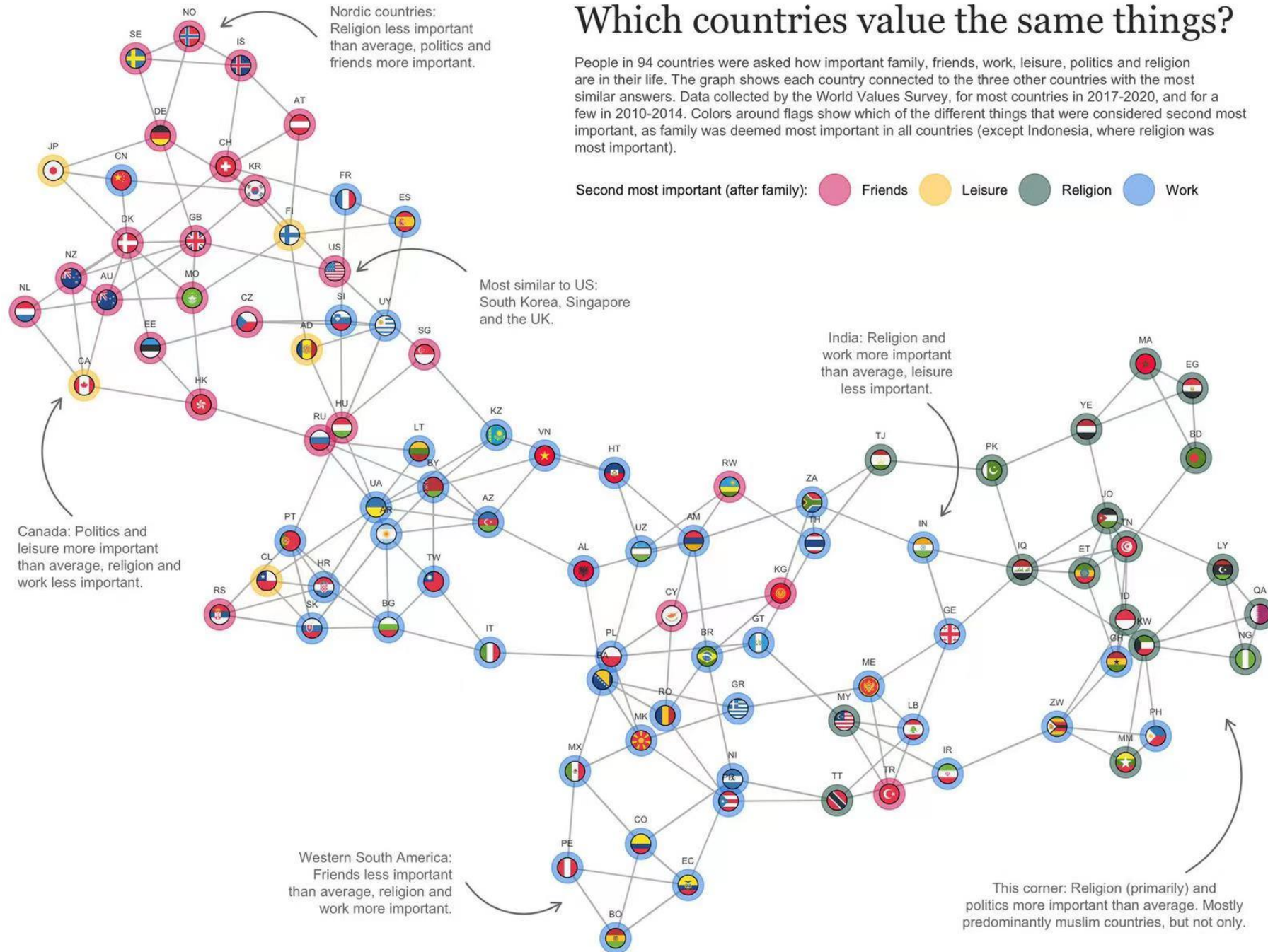


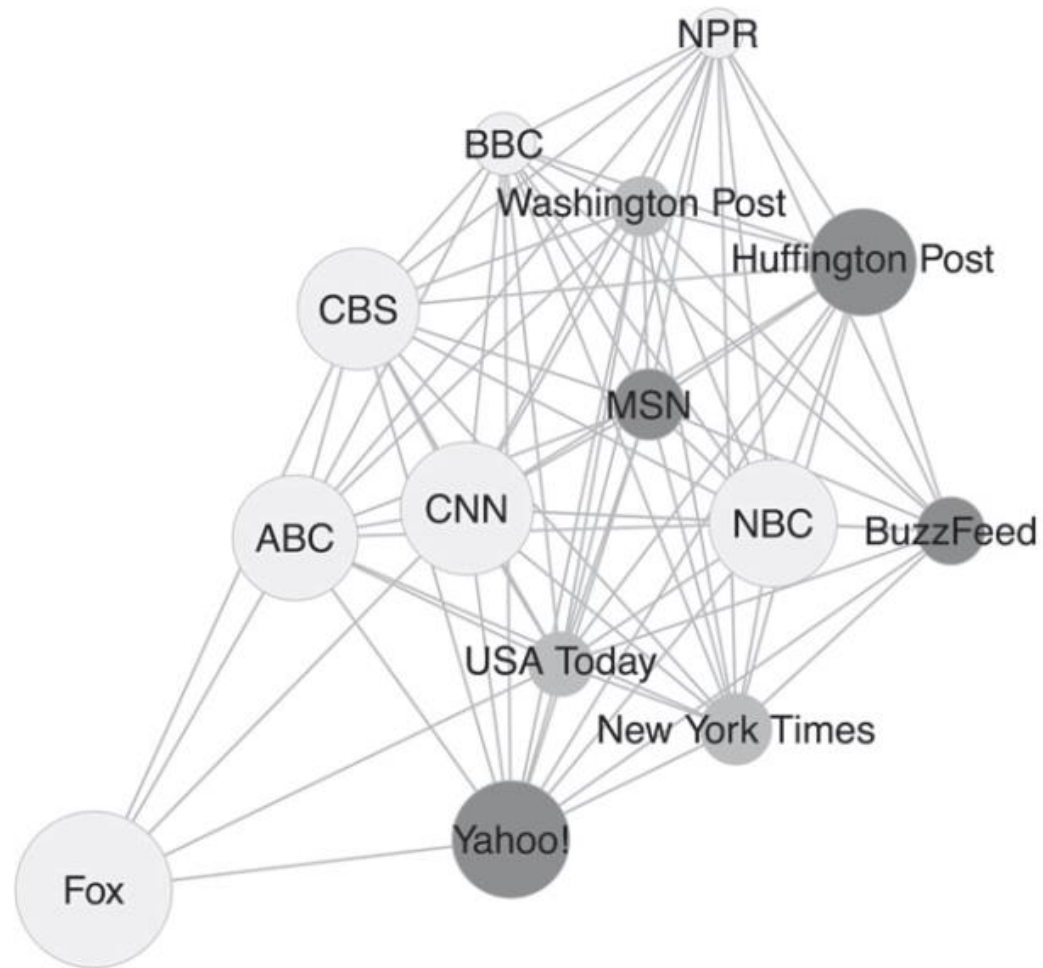
Co-authorship network of  
WKW School of Communication  
and Information, Nanyang  
Technological University (2014-  
2017)  
Data Source: Google Scholar

# Which countries value the same things?

People in 94 countries were asked how important family, friends, work, leisure, politics and religion are in their life. The graph shows each country connected to the three other countries with the most similar answers. Data collected by the World Values Survey, for most countries in 2017-2020, and for a few in 2010-2014. Colors around flags show which of the different things that were considered second most important, as family was deemed most important in all countries (except Indonesia, where religion was most important).

Second most important (after family): Friends (pink), Leisure (yellow), Religion (dark green), Work (blue)





**Figure 1** Example network showing cross-platform audience duplication in the United States.

Search Online

# Second-Hand Data

- Data collected by someone else
- Common sources of secondary data for social science include **censuses**
  - e.g. [Pew research center](#), comScore, Nielsen
- Open-sharing platforms
  - E.g., google dataset search
    - General Data Repositories
    - <https://datasetsearch.research.google.com/>
  - The Guardian
    - Crowdsourcing platform
    - <https://www.theguardian.com/politics/government-data>

# Google Dataset Search Beta

Search for Data Sets



Try [boston education data](#) or [weather site:noaa.gov](#)

Google Dataset search: <https://toolbox.google.com/datasetsearch>

[Kaggle](https://www.kaggle.com) (google): website  
for data scientist

Kaggle website interface showing the Datasets section. The header includes the Kaggle logo, a search bar, and navigation links: Competitions, Datasets, Kernels, Discussion, Learn, and a user profile icon. The Datasets section has buttons for Documentation and New Dataset.

The main content area displays a list of datasets, sorted by Hotness. The list includes filters for Public, Your Datasets, and Favorites. The search bar shows 10,486 Datasets. The list is sorted by Hotness.

| Rank | Dataset Name                              | Description  | Updated  | Version      | Tags  | Format    | Size     | Views                          |
|------|---|--|--|--------------|---|-----------|----------|--------------------------------|
| 30   | Google Play Store Apps                    | Web scraped data of 10k Play Store apps for analysing the Android market.  | Lavanya Gupta updated 13 days ago                            | (Version 1)  | video games, computer s..., internet, mobile web  | CSV       | 307.5 KB | 4 views, 1 comment, 3k likes   |
| 17   | Los Angeles Metro Bike Share Trip Data    | From Los Angeles Open Data   | City of Los Angeles Maintained by Kaggle updated 11 days ago | (Version 25) | socrata   | CSV, ODbL | 3 MB     | 6 views, 0 comment, 3k likes   |
| 21   | Annotated Honey Bee Images                | Apis mellifera across the USA with Location, Date, Health, and more labels | Jenny Yang updated 10 hours ago                              | (Version 2)  | animals, environment, natural res..., + 2 more... | CSV       | 50.1 MB  | 3 views, 1 comment, 2k likes   |
| 75   | LA Restaurant & Market Health Data        | From Los Angeles Open Data   | City of Los Angeles Maintained by Kaggle updated 11 days ago | (Version 22) | food and dr..., health, socrata                   | CSV       | 9.8 MB   | 12 views, 0 comment, 11k likes |
| 175  | Avocado Prices                            | Historical data on avocado prices and sales volume in multiple US markets  | Justin Kiggins updated 3 months ago                          | (Version 1)  | food and dr...                                    | CSV, ODbL | 628.7 KB | 48 views, 4 comment, 35k likes |
| 8    | SF Police Calls for Service and Incidents | From San Francisco Open Data   | City of San Francisco Maintained by Kaggle updated a day ago | (Version 69) | crime, socrata                                    | Other     | 163.7 MB | 3 views, 0 comment, 2k likes   |

Write an FOI Request - Your  
Right to Data



## Your Right to Data

- Know your rights
- Keep it focused
- Be specific and simple
- Submit multiple requests
- Ask for raw data

### Using FOI to Understand Spending

I've used FOI in couple of different ways to help cover COINS, the UK Government's biggest database of spending, budget and financial information. At the beginning of 2010, there was talk from George Osborne that if he became chancellor, he would release the COINS database to facilitate greater transparency in the Treasury. At this time it seemed a good idea to investigate the data in and structure of COINS so I sent a few FOI requests, one for the [schema of the database](#), one for the guidance Treasury workers receive when [they work with COINS](#), and one for the Treasury [contract with the database provider](#). All of which resulted in publication of useful data. I also requested all the spending codes in the database, [which was also published](#). All of this helped to understand COINS when George Osborne became chancellor in May 2010 and published COINS in June 2010. The COINS data was used in a number of websites encouraging the public to investigate the data including [OpenSpending.org](#) and the Guardian's [Coins Data Explorer](#).

After further investigation it seemed that a large part of the database was missing: the Whole of Government Accounts (WGA) which is 1,500 sets of accounts for public funded bodies. I used FOI to [request the 2008/09 WGA data](#) but to no avail. I also asked for the report from the audit office for WGA which I hoped would explain the reasons the WGA was not in a suitable state to be released. That was [also refused](#).

In December 2011, the WGA was released in the COINS data. However I wanted to make sure there was enough guidance to create the complete set of accounts for each of the 1,500 bodies included in the WGA exercise. This brings me on to the second way I used FOI: to ensure the data released under the UK transparency agenda is well-explained and contains what it should. I put in a FOI request for the [full set of accounts for every public body included in WGA](#).

*Lisa Evans, the Guardian*

Collecting first-hand data:  
Manually

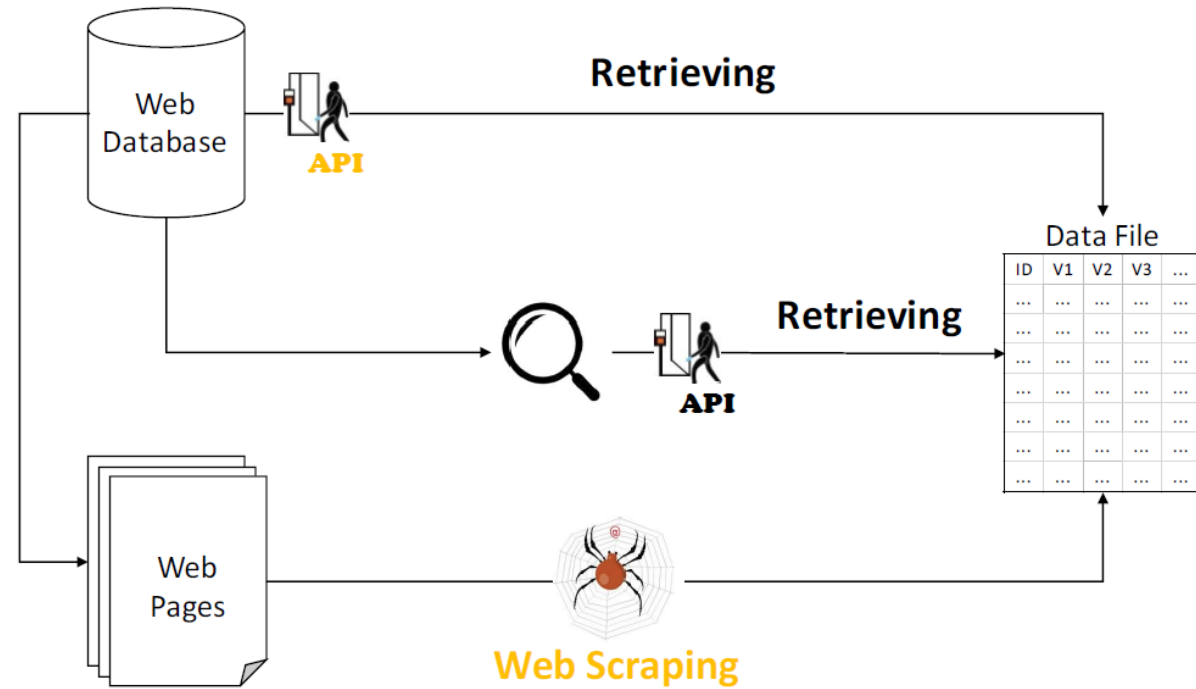
**Exercise I:**  
**Manually collecting the**  
**movie-actor network of**  
**marvel cinematic universe**  
Or recipe-ingredient network

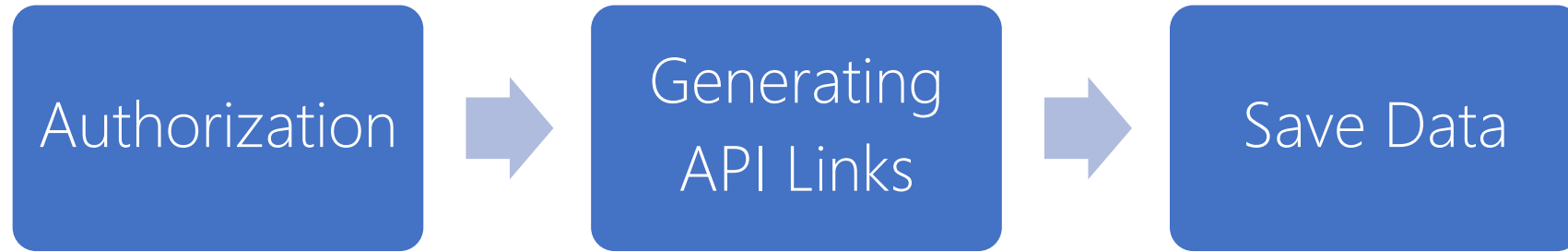


Collecting first-hand data:  
Through API

# Through API

- **API** (Application Programming Interface) is a small script file (i.e., program) written by users, following the rules specified by the web owner, to download data from its database (rather than webpages)
- An API script usually contains:
  - Login information (if required by the owner)
  - Name of the data source requested
  - Name of the fields (i.e., variables) requested
  - Range of date/time
  - Other information requested
  - Format of the output data
  - etc.





## Collecting through API

- Authorization (Log in): Many social media platforms require a formal procedure of authorization.
  - Create an application on the website
  - Authorize the app to access data using a protocol called OAuth
  - Four credentials: Consumer key, consumer secret, access token, and access token secret
- Access data through API link (<https://developer.twitter.com/en/products/twitter-api>)

# Twitter API

Use the Twitter API to Listen to and analyze the public conversation

| Tweets                      | Standard (free) | Premium | Enterprise |
|-----------------------------|-----------------|---------|------------|
| Publish and engage          | ✓               |         |            |
| Search Tweets: 7-days       | ✓               |         |            |
| Search Tweets: 30-days      |                 | ✓       | ✓          |
| Search Tweets: Full-archive |                 | ✓       | ✓          |
| Filter Tweets               | ✓               |         | ✓          |
| Sample Tweets               | ✓               |         | ✓          |
| Batch Tweets                |                 |         | ✓          |
| Direct Messages             | ✓               |         |            |
| Account and users           | ✓               | ✓       | ✓          |
| Metrics                     |                 |         | ✓          |
| Ads API                     | ✓               |         |            |
| Publisher tools and SDKs    | ✓               |         |            |

## Collecting through API (twitter)

1. Behavioral data: the distribution and frequency of content on social media
2. Behavioral data: information-sharing behavior, number of retweet especially
3. Behavioral data: time-stamp
4. Network data: Friendship network online
5. Network data: Information-sharing network
6. Content data: text, what are they talk about
7. Content data: text, emotions in the text




# nyt API

{ } Developers


HomeAPIsCovid-19 DataGet StartedSign In

APIs


Filter by title & description




**Archive API**  
Get all NYT article metadata for a given month.




**Article Search API**  
Search for New York Times articles.




**Books API**  
Get NYT Best Sellers Lists and lookup book reviews.




**Community API**  
Get user comments. (BETA)




**Geo API**



**Most Popular API**



**Movie Reviews API**

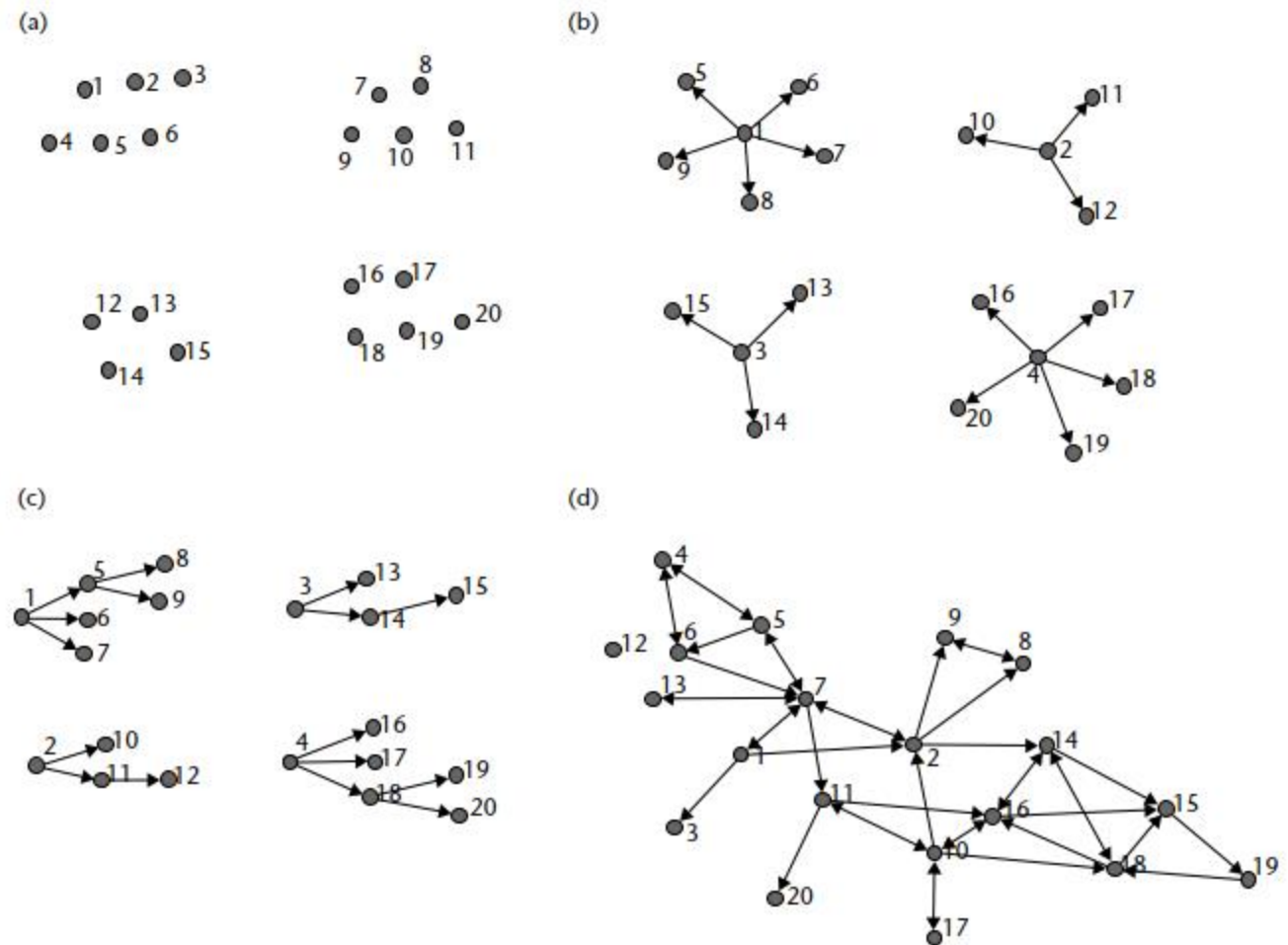


**RSS Feeds**

Collecting first-hand data:  
Survey and Questionnaire

# survey sampling techniques

- a) random selection (although clustered)
- b) Egocentric
- c) sequenced or snowball
- d) census



**Figure 3-1.** An illustration of four different survey sampling techniques: (a) random selection (although clustered), (b) egocentric, (c) sequenced or snowball, and (d) census.

# Survey

- Survey network data consist of asking individuals whether they talked to or consulted anyone about some topic (name generator).
- Question format:
  - If you ask people to *recall* names (an open list format), fatigue will result in underreporting
  - If you ask people to check off names from a full list, you can often get over-reporting
- It is common to limit people to ~5 nominations. This will bias network stats for stars, but is sometimes the best choice to avoid fatigue.
- Concrete relational indicators are best (who did you talk to?) over attitudes that are harder to define (who do you like?)

# Survey sampling techniques

| Techniques    | Descriptions   | Example  |
|---------------|--|--|
| 1. Survey     | Standard survey questions  | “Have you consult anyone about vaccination?”   |
| 2. Egocentric | Name generators and questions on the interaction between those named | Who are your close friends in university (names 3-5), do your friends know each other? |
| 3.Sequenced   | Snowball: Index cases name alters and all alters are interviewed     | Close contacts in pandemic   |
| 4.Census      | Roster: All members of a community are interviewed                   | Marriage   |

## **Exercise II:**

# **Applying Twitter API**

# Exercise III: Collecting Census data through roster

- name your friends by selecting from a class roster.

