Analyzing Wikipedia collaboration networks
(background)

Jürgen Lerner

University of Konstanz

Winter 2016/2017
(last updated: October 26, 2016)
Outline.

Wikipedia data.

Experimental data vs. observational data.

Prediction vs. explanation.
Outline.

Wikipedia data.

Experimental data vs. observational data.

Prediction vs. explanation.
Wikipedia: data, structure, variables.
Core structure: network connecting users with articles.

A lot of additional information attached to users, articles, links, 
...
General Wikipedia data sources.

Information about Wikipedia database download

Database schema (table names and fields)
https://www.mediawiki.org/wiki/Manual:Database_layout

Finding the actual files https://dumps.wikimedia.org
- click on Database backup dumps
- select the wiki of interest (e.g., enwiki)
- if the dump is in progress or failed: Last dumped on . . .
Wikipedia database dumps: individual files (1/2).

The SQL files (ending with sql.gz; at the bottom of the page) each encode the current state of one table.

Important SQL files
- page.sql.gz (id, namespace, title, is_redirect, length)
- {page|external|category|...}links.sql.gz links of various types
- user_groups.sql.gz assignment of users to groups (special rights)
The XML files encode revisions of pages (some the **current state**, some the **whole history**).


- **stub-meta-history.xml.gz**
  
  whole page history; except the pages’ text

- **pages-meta-current.xml.bz2**
  
  only the last revision, including text

- **pages-meta-historyX.xml-pYpZ.7z**
  
  whole page history, including text
Downloading the history of single (or few) pages.

Up to 1000 revisions can be downloaded from

**WikiEvent**: download the complete history of few pages:
http://algo.uni-konstanz.de/software/wikievent/

**Wikimedia API** can be used for download and editing:
https://www.mediawiki.org/wiki/API:Main_page
Quality of Wikipedia articles.

Articles are assigned one of the following assessment grades:

- Featured: 4,749
- A-Class: 1,563
- Good articles: 25,634
- B-Class: 106,402
- C-Class: 218,506
- Start: 1,373,326
- Stub: 2,818,537
- Unassessed: 501,830

Assessment is done by the community (social evaluation).

**Homework** (until next week): find out how to determine the assessment class of all articles.

https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment
Featured article assessment procedure.

- Users can **nominate** *featured article candidates*.¹
- Users can discuss, **support**, **oppose**, these nominations.
- Articles are promoted if there is (positive) **consensus**.

Featured article candidate nomination, discussion, and voting are fully transparent.

- Featured articles can be **demoted** if they no longer meet the criteria.

Elections for administrator status.

Discussion and voting for candidates:
Outline.

Wikipedia data.

Experimental data vs. observational data.

Prediction vs. explanation.
Empirical research: experimental data vs. observational data.
Empirical research seeks to find relations among variables in data, such as “the more $X$, the more $Y$.”

Thus, it needs data in which $X$ shows some variation.

In an experimental setting, this variation is produced deliberately and systematically.
Empirical research: experimental data.

In an **experimental** setting, variation in variables is produced deliberately and systematically.

Example: assessing the effect of a (pharmaceutical) drug

▶ sample instances from a population and randomly divide into two subsets;
▶ subset one gets the drug; subset two not (or placebo);
▶ measure the variable of interest (e.g., health) and compare the distribution in the two subsets.

Advantage: randomization should equalize the distribution of any confounding variable (known or not known).

Disadvantage: some variables are hard to vary experimentally (e.g., number of friends; team diversity); ethical considerations.
Empirical research: observational data.

Empirical research seeks to find relations among variables in data, such as “the more $X$, the more $Y$.”

Thus, it needs data in which $X$ shows some variation.

When analyzing observational data, this variation has to be present in the data; results from some unknown mechanism.
Empirical research: observational data.

When analyzing **observational** data, variation in variables has to be present in the data; results from some unknown mechanism.

Advantage: data is realistic; analyzing whole population is possible.

Disadvantage: unknown variables might confound the effects of interest. Direction of causality hard to establish.

Danger: insights like “hospitals increase the mortality.”

Remedy: open data and methods; replication, replication, . . .
Outline.

Wikipedia data.

Experimental data vs. observational data.

Prediction vs. explanation.
Empirical research: prediction vs. explanation.
Different goals: prediction vs. explanation.

**Prediction** wants to predict the data yet to come.

Example: numerical weather forecast.

**Explanation** wants to explain why something happens.

Example: Navier-Stokes equation

$$\frac{\partial u}{\partial t} + u \cdot \nabla u = -\frac{1}{\rho} \nabla p + \nu \nabla^2 u + \frac{1}{3} \nu \nabla (\nabla \cdot u) + g$$

\[\text{http://celebrating200years.noaa.gov}\]
Another theory seeking explanation.

**Evolutionary theory** explains that living species change as a function of mutation, selection, and inheritance.

It is quite useless for predicting how species will develop in the future.
Prediction vs. explanation.

The distinction is important but the question which of the two goals is better, or more scientific, is nonsense (in my opinion).

Additional background reading:

Compares algorithmic modeling with data modeling (rather than prediction vs. explanation).