Analyzing Wikipedia collaboration networks
(introduction)

Jürgen Lerner

University of Konstanz

Winter 2016/2017
(last updated: October 26, 2016)
Wikipedia: encyclopedic articles written by volunteers.

How do Wikipedians organize themselves?
Context discipline: computational social science.

Log data from novel ways to communicate, interact, collaborate, search, purchase, yield comprehensive, fine-grained data about human behavior, facilitating the emergence of computational social science.¹

This field is naturally interdisciplinary, involving at least social science and computer science.

Why studying Wikipedia?

Wikipedia is an organization in which actors are creating products that are intended to have appeal to others.

Facilitates the analysis of large task-oriented groups.

Studying Wikipedia has some outstanding positive aspects:

- relevance of the created products (people are using them);
- novel form of organization (self-governance, volunteers);
- large-scale, fine-grained, and complete data availability;
- clearly defined intended outcome & indicators for success.

⇒ Studying Wikipedia is relevant and convenient.
What are we doing in this seminar?
Outline of this seminar.

Read and discuss published empirical research that analyzes Wikipedia data (eight different papers).

- identify the paper’s **research question(s)**;
- understand the **analysis method**;
- understand the relevant **variables** and how they are obtained/computed from the data (precisely!);
- check what the **results** tell us about the question;
- **critically discuss** the paper:
  - are there alternative explanations for the results?
  - what are the limitations or faults of the paper?
  - how could we improve the analysis; the definition/computation of variables?

Learn about state-of-the-art approaches to analyze large-scale collaboration networks.
Structure of this seminar / deliverables.

Each participant chooses one paper as a presenter and main discussant;

one paper is presented per week (talk of approx. 30 minutes).

The presenter should discuss the talk with me before the presentation; hands in the slides on the day of presentation (will be posted on the website).

Every participant reads every paper before it is presented and actively participates in the discussion following the presentation.

The presenter writes a term paper (approx. 6–10 pages) about the one chosen paper, covering the points from the previous slide (including issues raised in the discussion), due at the end of the semester.
IMPORTANT: To-do until next week.

Topic assignment:

- Participants send me per email **until Monday 31 October**
  a preference list of the topics:
  1. most preferred topic
     ...
  8. least preferred topic

- I select randomly from those most interested in a topic and post these assignments on the Webpage before the seminar next week.

If you want to participate but have no preferences, send an email without the list.

In any case, register for the course in the Studis/LSF!
Topics.

Hypothesize that the quality of Wikipedia articles will be

- lower if there is conflict in the team;
- higher if team members’ orientation is towards content production.

Furthermore, they hypothesize that a team’s administrative orientation decreases conflict and that conflict moderates the effect of diversity: if conflict is high then diversity has a positive effect on quality.
(2) Article quality by membership turnover.


Hypothesize that membership turnover relates to article quality in a non-monotonic fashion: increasing quality up to an optimum level of turnover and decreasing it thereafter.
(3) Promotion to administrator status.


Analyzes indicators for past user behavior that increase or decrease the probability of being elected as administrator; compares to promotion criteria stated on Wikipedia.

Analyzes the influence of properties of the candidate and voter on the probability for voting positively. Puts emphasis on relative assessment and on the dependence on previous votes.

Analyze whether voting behavior follows the predictions of structural balance theory or status theory.

(Ignore the results from the “Epinions” and “Slashdot” data.)
(6) Reputation of Wikipedia users.


Develop a method to compute the reputation of users based on the stability of their contributed text and of their edits. Test the predictive value of their reputation indicator.
(7) Analyzing the article-user network.


Analyze what draws specific users to contribute to specific articles.
(8) Coordination costs and benefits.


Analyze patterns in discussion ("coordination") in Wikipedia. Distinguish discussion associated with high-quality articles and average articles.

(Ignore the results from the GitHub data.)
Tentative schedule.
Tentative schedule.

Until Monday 31 October: send topic preference-list to me.

Next one/two weeks: additional background information, including where to obtain Wikipedia data.

Next three weeks: read the two references listed under “background reading.”

Presentations start on 16 November (one presentation per week; in the given order; see Webpage).