

Social Network Analysis for Computer Scientists

Fall 2026 — Assignment 2

<https://liacs.leidenuniv.nl/~takesfw/SNACS>

Deadline: November 8, 2026 (Anywhere on Earth)

This document contains two exercises that each consist of various numbered questions that together form Assignment 2 of the Social Network Analysis for Computer Scientists course taught at Leiden University.

For each question, the number of points awarded for a 100% correct answer is listed between parentheses. In total, you can obtain 100 points and 10 bonus points. Your assignment grade is computed by dividing your number of points by 10. Please do not be late with handing in your work. You have to hand in the solutions to these exercises **individually** or in a **team of exactly two students**. Discussing the harder questions with fellow student(team)s is allowed, but writing down identical solutions is not.

Clearly and concisely describe how you obtained each answer. Write down any nontrivial assumptions that you make. Always clearly describe which tools and packages you used. Recall that omission of a reference to source material is considered plagiarism. Include relevant source code in an Appendix that you reference in your answers. Please use the `listings` package for including source code. If you used an interactive notebook, then there are two options: 1) use `nbconvert` to convert the notebook to regular source code and make sure that you clearly delineate, for example using comments in the code, which code pertains to which question, or 2) attach your interactive notebook as a readable PDF attachment to your submission. In both cases, the code is an Appendix of your report, and your answers to the questions must be in the report.

Submission. Hand in your report via Brightspace, in one `.pdf` file, typeset using $\text{\LaTeX}$. Remember to specify your name and student ID (ULCN number) on top of your assignment.

Do not copy the full text of each question into your document (if you do, you will be asked to resubmit). Just stating the question number and then your answer, is sufficient. If you really need to submit multiple files, please attach them all in one submission. If you want to make a new submission, replacing your previous submission, make sure to again include all the files in that submission.

Questions or remarks? Ask your questions during one of the weekly lectures or lab sessions. Good luck!

Exercise 1: Twitter mention network analysis (45p + 10p bonus)

Two samples of Twitter datasets can be found at

<https://liacs.leidenuniv.nl/~takesfw/SNACS/twitter-small.tsv> and
<https://liacs.leidenuniv.nl/~takesfw/SNACS/twitter-larger.tsv>.

The full dataset (not required until bonus Question 1.6) can be found in the LIACS REL environment (see <https://rel.liacs.nl>, only accessible internally or via LIACS SSH/VPN):

`/data/SNACS/twitter.tsv`.

Up until now, we have only looked at social network data which was already in a nicely formatted edge list. In practical network analysis research, this rarely ever happens. Therefore we will now work with real raw data from a Twitter crawl [1]. The dataset `twitter.tsv` contains over 450,000,000 tweets, crawled from June 2009 to December 2009. The file `twitter-small.tsv` contains a small subset of these tweets that can, after preprocessing, be handled with Gephi, whereas `twitter-larger.tsv` contains a bit larger subset that can be analyzed using for example NetworkX. Each line of these files contains one tweet, consisting of three tab-separated (`\t`) fields denoting the timestamp, user who sent the tweet and the content of the tweet. For example:

```
2009-07-05 14:07:18    aeneas    Hi @achilles, how are you? #old
```

In the tweet content, a word starting with the `@` symbol (such as `@achilles`) means that user `achilles` is being mentioned by user `aeneas`, indicating that the tweet by `aeneas` was directed at or specifically about `achilles`. We refer to this as a *mention*. Mentions are the most direct sign of public communication on Twitter. Tweets can also be directed at more than one user.

The *mention graph* is a Twitter network represented by a directed graph. The set of nodes consists of users (anyone sending out a tweet or being mentioned by someone else in a tweet). The set of links consists of all user pairs (x, y) such that user x mentioned user y at least once. The mention graph is in fact a *weighted* directed graph where the number of times a user x mentions another user y is the link weight. Now, for the `twitter-small.tsv` dataset, answer Questions 1.1–1.4.

(12p) Question 1.1 Extract the mention graph from the Twitter data. Relevant steps could be:

- Parse the input file line by line (for example using Python or Perl).
- Generate the adjacency list: for each user (identified by its username), keep a list of the users that this user mentions, and count the number of mentions.
- Output the adjacency list as a weighted edge list `tsv`-file.

Discuss the steps that you took, and describe the issues that you ran into while parsing this “real-world” data, and how you solved them. For example, discuss possible text mining and parsing issues. Carefully think of how to capture a valid Twitter username. Important: from your description (in words; do not just give the code), it should be possible to unambiguously reproduce your network dataset; points are mostly awarded for reproducibility of your approach from your description, and not merely for code correctness.

(12p) Question 1.2 Present relevant statistics of your mention graph in a table, including at least

- (1p) the number of nodes and edges,
- (1p) number of strongly and weakly connected components,
- (1p) number of nodes and edges in the largest of these components,
- (1p) density,
- (1p) (approximated) average node clustering coefficient, and
- (1p) (undirected) (approximated) average distance in the giant component.

Moreover, generate the following diagrams and include them in your report as figures:

- (3p) indegree and outdegree distributions, and
 - (3p) (undirected) (approximated) distance distribution of the giant component.
- (4p) **Question 1.3** A weight distribution indicates how often each link weight occurs. Present this distribution using a proper diagram. What type of distribution does this appear to be?
- (5p) **Question 1.4** Determine the top 20 users based on three different centrality measures (for example, betweenness, closeness and degree centrality). Mention how you deal with directionality. Discuss the results.
- (12p) **Question 1.5** Answer Question 1.2 for the larger dataset `twitter-larger.tsv`.
- (10p, bonus) **Question 1.6** Answer Question 1.2 for the full 450M tweet dataset given in `twitter.tsv`. This is very challenging, and may require you to either use large computing resources, or to be smart/selective in your data processing steps, for example by systematically filtering certain users and/or links, e.g., based on some threshold for the number of mentions. It is recommended to only do this after answering all the other questions in this assignment. Again, remember to explain your approach to dealing with the sheer size of this dataset.

[1] J. Yang and J. Leskovec, Temporal variation in online media, in *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pp. 177–186, 2011. Available at <https://doi.org/10.1145/1935826.1935863>

Exercise 2: Evolution of a Social Network of Hamsters (55p)

Hamsterster was an online social networking website for hamster owners, where users created profiles representing their pet hamsters rather than themselves. Profiles could include photos, personal information, and characteristics of the hamsters, similar to social networking platforms for humans. The network existed in the period 2004–2011, and in this assignment we aim to understand the structure and evolution of this network over time.

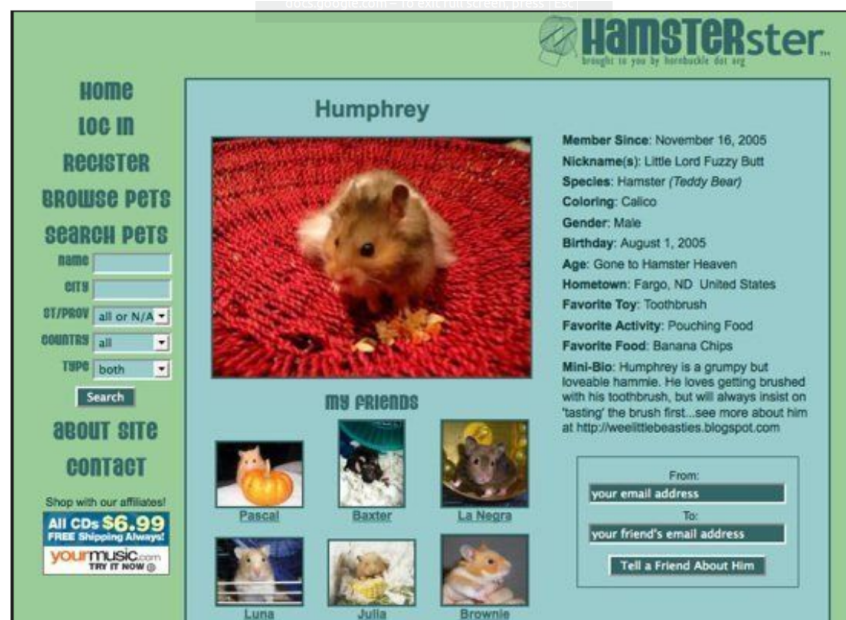


Figure 1: The Hamsterster social networking website.

A dataset of this social network [2], in terms of a list of undirected edges and corresponding list of node attributes, can be found in the following two files:

```
https://liacs.leidenuniv.nl/~takesfw/SNACS/hamster-edges.tsv  
https://liacs.leidenuniv.nl/~takesfw/SNACS/hamster-nodes.tsv
```

The edges do not have timestamps. Assume that an edge exists from the year in which both of its endpoints have joined the network, i.e., from the later of the two `joined_year` values of its endpoints.

Answer the following questions about this social network data, using results obtained through NetworkX or another package for network analysis. It is recommended to use Gephi for an initial visualization.

- (15p) **Question 2.1** Present relevant statistics on the *evolution of the structure* of this social network over time, including at least basic size counts, degree distribution, clustering and average distance over time. For example, plot these values and distributions for 8 snapshots of the network including all data up to 2004, 2005, . . . , 2011. Briefly interpret your results based on what is known in the field about the (evolution of the) structure of real-world networks.
- (10p) **Question 2.2** Generate networks according to two relevant *synthetic network models* discussed in the course, and compare the resulting network structure with that of the structure of the 2011 version of the Hamsterster network. For each of the two models, describe and interpret the differences in structure.
- (5p) **Question 2.3** Investigate *attribute assortativity* in this network. Present your results in a meaningful way, such that you can interpret the results.
- (15p) **Question 2.4** Apply two different *community detection* algorithms (e.g., modularity maximization and degree-corrected SBM) to this dataset. Describe your approach (parameters, etc.). To what extent do the communities agree with a partitioning of the network into communities based on a particular attribute of the nodes? Use a relevant representation (table or diagram) to present the result of this comparison, and provide a short interpretation.
- (10p) **Question 2.5** Imagine that this social network of hamster owners still exists today, and consists of too many nodes and edges to analyze at once. Create two different samples, using two of the different *sampling algorithms* discussed in the course, of this social network dataset, at a fixed sampling rate (e.g., 25%). Explain what the idea behind each sampling approach means in this particular social network. Compare the resulting networks with the original network in terms of at least the degree distribution and clustering. Interpret how the resulting samples deviate from the full network dataset.

[2] R.A. Rossi and N.K. Ahmed, The network data repository with interactive graph analytics and visualization, in *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 4292–4293, 2015. Available at <https://networkrepository.com>