University of Zurich[UZH]

Department of Informatics

University of Zürich
Binzmühlestrasse 14
CH-8050 Zürich
Switzerland

Prof. Dr. Burkhard Stiller
Andri Lareida
Thomas Bocek
Communication Systems Group
CSG@IFI
Phone: +41 44 635 6751(Andri)
Fax: +41 44 635 6809
E-mail: stiller@ifi.uzh.ch
lareida@ifi.uzh.ch
URL: http://www.csg.uzh.ch/

Assignment (Vertiefungsarbeit) for Robin Stohler

Task Description:	Andri Lareida, Dr. Thomas Bocek
Title:	**Scaling Social Network Analysis**
Start Date:	30 October, 2015
End Date:	30 January, 2016
Supervisor:	Andri Lareida, Thomas Bocek
Location:	Home
Support:	VIOLA, Hadoop

1. Introduction and Motivation

Social Network Analysis (SNA) traditionally investigates individuals and their social relations. Individuals and relations are modelled as a graph consisting of nodes (individuals) and edges (relation, e.g., friendship). Modelling a social network as a graph enables the calculation and interpretation of several graph metrics, such as Centrality Metrics, Symmetry Measures, and Entropy Measures [3]. Libraries for the calculation of these measures are available for platforms [5] and can, therefore, be easily applied to an accordingly prepared data set. However, the interpretation of these measures strongly depends on the mapping chosen to model an observation as a graph. SNA methods are not limited to social networks [2], in fact, these can be applied to any network that can be modelled by nodes and edges. The Orgnet definition of SNA: “Social network analysis [SNA] is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities.” [4] gives hints on the applicability of SNA to areas beyond Social Networks.

In Order to apply SNA metrics to the BitTorrent network [1] this assignment shall investigate the possibilities to scale these graph calculations to very large graphs. Very large meaning in the order of Millions of nodes and Billions of edges. The goal is to identify viable options of calculating SNA metrics, especially clustering for such a graph in meaningful time.

2. Description of Work

The Student will perform research on the topic of scalable calculation of network metrics and experiment with different tools if required. The work consists the following points:

- Research on tools for network metric calculation, especially algorithms / libraries for Hadoop
- Comparison of different tools
- Comparison with sample data if require, *i.e.,* if no performance data is available
- Documentation of findings in written report

3. Thesis Goals

The VA shall answer the following questions:

- What options are available to calculate SNA metrics?
- How do these options scale to large graphs, *e.g.*, Millions of nodes and Billions of edges
- What are the requirements in terms of infrastructure?
- Documentation of the findings in a Report.

4. Activities

Based on the description of work, the following tasks targeting the required milestones need to be accomplished:

- Milestone 1: Planning
  Present a schedule detailing the work plan in the 1st Week.
- Milestone 2: Available tools and libraries investigated.
- Milestone 3: Final Presentation and Demo of Prototype
  Working example is presented.

5. General Notes

- The student has to provide a written schedule for his/her full thesis steps within the first two weeks of his/her work. Clarify details with your supervisor and finalize the schedule of tasks, basically in a weekly fashion. Include the final presentation, too.
- At the assignment's end (date to be set) a final public and self-containing presentation and demonstration (20 min. max plus Q&A) has to be given.
- Establish periodic meetings with the supervisor to report on progress and to discuss problems.
- Students involved in this thesis are required to read and answer e-mails related to this project at least three times a week. A more frequent, if needed, interaction will determine a better basis for supervision and progress.

The information sheet on important hints for thesis work is required to be known.

6. Formal Results

Besides the final oral presentation of your work the following written documents are part of your work and need to be handed in to the supervisor in time:

- A report in a soft cover binding and in 3 copies (in English, preferred): It covers the problem to be solved, the discussion of the design choices, a set of arguments on the final design choice, a list of solved and open issues, a table of content and figures (including tables), a valid list of bibliographic references, and optional appendices as required. A critical consideration of the task, the assignment, and the result will conclude the report. The official acknowledgement section is mandatory, a personal one optional, however recommended, as usually a number of people took part

in the process of finalizing the thesis. The text processing can be done in FrameMaker or LaTeX.

- The student can choose between writing a report as outlined just here above or taking an oral exam of approximately 25 minutes about all topics covered in and besides this assignment. The oral exam starts with a 20 min presentation of the candidate, where the slide set has to be handed in, including annotations and remarks in comments fields, seven days prior to the exam date to be set.
- A dedicated CD has to be produced containing: the written thesis (report) in source files, figures in source file and gif or eps, and a full printable PDF file, the set of slides for the final presentation in source, and all further material used, if available in electronic form, such as all existing and documented code, scripts, scenarios, plans, and results.
- A German (or English, in case of a German report) summary of maximum 1 page, which will enable a quick and clear survey of this assignment's tasks and results. This summary will be part of the bound report, and is included after the front page and before any other text will follow. It includes content-wise four areas of concern: introduction, aims and goals, results, and further work.
- The complete set of copies of the report, the CD, and the talk must be completed and handed in to the supervisor in time before the assignment will change into "submitted" status. After that a date of the oral exam will be fixed between the responsible professor and the candidate. One of the supervisors should be part of the oral exam as well.

7. References

The following list of references addresses key aspects and serves as a starting point for the work. Further papers, scenarios, and document research is a must! It is more than highly recommended and may be essential for a successful completion of the work:

[1] "BitTorrent.org". http://www.bittorrent.org/index.html

[2] Burger, Valentin, et al. "Social Network Analysis in the Enterprise: Challenges and Opportunities." Socioinformatics-The Social Impact of Interactions between Humans and IT. Springer International Publishing, 2014. 95-120.

[3] Hoßfeld, Tobias, et al. "On the computation of entropy production in stationary social networks." Social Network Analysis and Mining 4.1 (2014): 1-19.

[4] Orgnet, "Social Network Analysis, A Brief Introduction". http://www.orgnet.com/sna.html

[5] The igraph core team, "igraph". http://igraph.org/

Zürich, October 30, 2015

Prof. Dr. Burkhard Stiller