Web Science

Project Description

Julia Neidhardt

julia.neidhardt@ec.tuwien.ac.at

Vienna University of Technology
Institute of Software Technology and Interactive Systems
E-Commerce Group
Favoritenstrasse 9-11/188/4
A-1040 Vienna, Austria

Project – Assignment

- **Extracting and Analyzing Network Data**

In this project, students will collect and analyse network data with R. The goal is to discuss major structural and dynamic points of the collected networks as well as to visualize them.

30 Points maximum

Project – Instructions

1. **Group Formation**
    - Form groups of 2 persons (by December 13th, 2016)

2. **Collection of network data**

Capture a series of Twitter data samples, which are all related to one specific topic that you are free to choose. Examples:

- Partial networks (related to one hash-tag or search term)
- Ego networks (follower relationships focusing on important users in partial network)
- Longitudinal data (partial network at different points in time)

Project – Instructions (2)

3. Data Analysis and visualization (10 points)

- **Network statistics** Calculate different metrics and measures, e.g., number of vertices, number of edges and their weights, degree distribution (resp. in-degree and out-degree in the directed case), centrality indices, clustering coefficient, reciprocity, shortest paths, network diameter, density, number and size of connected components.

- **Communities** Identify communities/groups and their structure.

- **Visualization** Use different layouts to find and highlight patterns and relationships. To make sense of the data more easily, map the metrics and measures calculated onto visual properties such as size, colour, and opacity.

Project – Instructions (3)

- **Longitudinal data (5 points)**
  - How did the network change? What dynamics can be identified? Are there critical events in the period observed?

- **Text analysis (5 points)**
  - What do the users talk about? Are there any keywords/topics that are particularly important? Does this vary in different groups/communities? Are the discussions controversial?

Project – Instructions (4)

- **Discussion and conclusion (5 points)**
  - What about the quality of the data? What are the limitations? What questions can be answered and what new insights can be gained by analyzing the structure of the networks? What nodes are in strategic locations? What can you tell about the communities identified? What do you learn by looking at the different networks? Why is it interesting to combine the information from structural and textual data? Interpret and discuss your findings!

Project – Instructions (5)

For each group:

- Import and analyse a Twitter partial network (search for a keyword or hash-tag) and analyse the re-tweet relations. In addition, look also at the contents of the tweets and integrate insights and conclusions into your analysis.

- To study longitudinal effects collect this network at several points in time and compare relevant properties of these networks (it will depend on the selected topic, which time differences for data collections make sense).

- Select two interesting users and analyse their follower networks (also consider the followers' followers or some of them if there are too many). Compare the networks.

Project – Instructions (6)

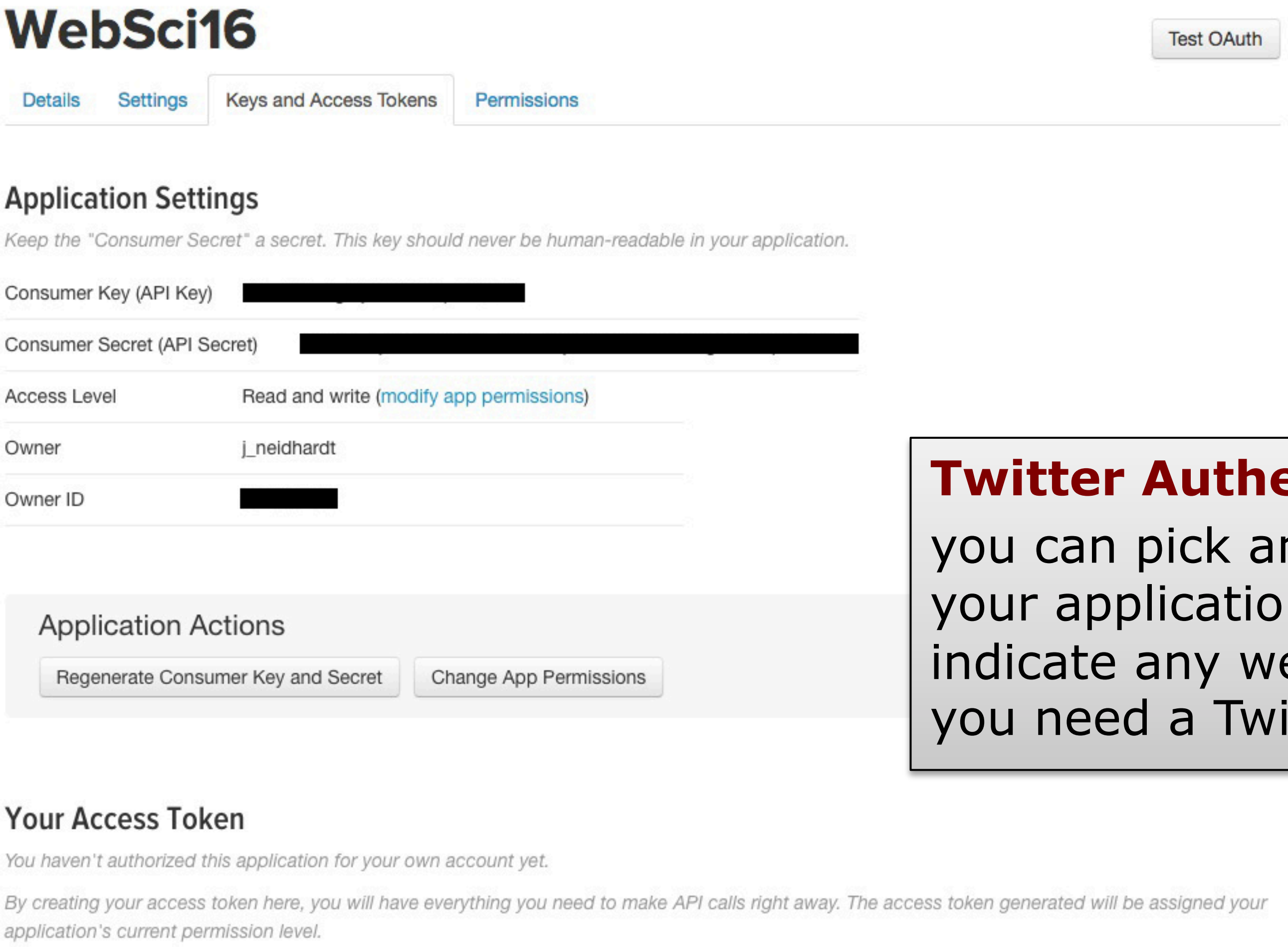
- **Report (25 points maximum)**
  - Compile a report to document and summarize your work. It should contain a title page: **title**, **names** and **abstract** (150-200 words).
  - Upload the **report** (pdf) as well as an **R file**, which contains the code of your analysis, and the **data** that you gathered to TUWEL **by January 16th, 2017**

- **Presentation (5 points maximum)**
  - Prepare a short presentation (12-15 minutes, 10-12 slides maximum)
  - Upload the presentation (pdf) to TUWEL **by January 16th, 2017**
  - Presentations will take place **on January 17th, 2017** (Attendance required!)

Project – Remarks

- Plenty of online resources are available that show how Twitter data can be gathered and analysed with R. Here some of them are listed:
  - Collections of blogs related to R and Twitter: https://www.r-bloggers.com/search/twitter/
  - Twitter authentication with R: http://thinktostart.com/twitter-authentification-with-r/
  - Obtaining a Retweet network: http://thinktostart.com/visualize-retweets-with-r/
  - Construction of Follower network: https://www.r-bloggers.com/graphing-twitter-friendsfollowers-with-r-updated-yet-again/
  - Comprehensive overview of Twitter content and structure analysis: http://wombat2016.org/slides/yanchang.pdf

Project – Remarks (2)



Application Settings

Keep the "Consumer Secret" a secret. This key should never be human-readable in your application.

Consumer Key (API Key)	[redacted]
Consumer Secret (API Secret)	[redacted]
Access Level	Read and write (modify app permissions)
Owner	j_neidhardt
Owner ID	[redacted]

Application Actions

Your Access Token

You haven't authorized this application for your own account yet.

By creating your access token here, you will have everything you need to make API calls right away. The access token generated will be assigned your application's current permission level.

Twitter Authentication: you can pick any name for your application and indicate any website (but you need a Twitter account)

Project – Remarks (3)

- **Data gathering**: try to start as soon as possible to collect data for the project as you might want to try different search terms / hash-tags / number of followers etc. – Twitter has a rate limit, so all this might take a while

- **Manipulating the data**: familiarize yourself with R, you will need to access lists and dataframes. A short introduction is provided here http://kateto.net/networks-r-igraph

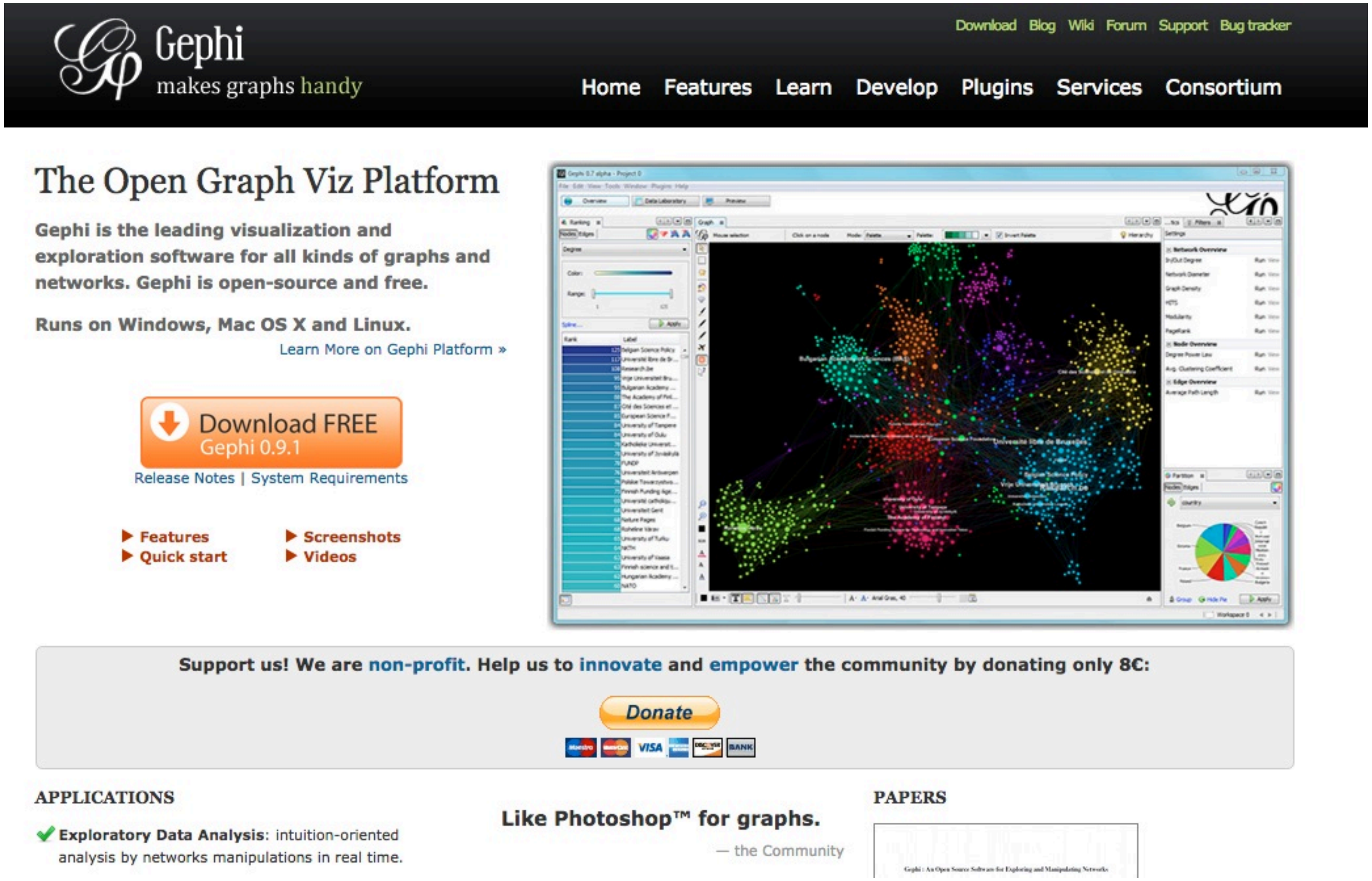
- **Weighted Re-tweet network** (based on http://thinktostart.com/visualize-retweets-with-r/)
  - Add *E(rt_graph)$weight <- 1*
  - *rt_graph <- simplify(rt_graph, edge.attr.comb=list(weight="sum"))* after *rt_graph = graph.edgelist(retweeter_poster)*
  - Check weight distribution: *table(E(rt_graph)$weight)*

Project – Remarks (4)

- **Text analysis**: take a close look at the text of the tweets to better understand the dynamics of the discussions. This complements the structural analysis. You can also try to apply some basic text mining to identify frequent terms (but you don't have to). How this is done is for example explained here: http://wombat2016.org/slides/yanchang.pdf

- **Other social network analysis tools**: You can export the graph data in a graph file format (e.g., graphml, gml, etc.) or as csv file in order to access the data with other analysis tools. For example, if you want to do nice visualizations the tool Gephi is recommended.

Gephi (http://gephi.org)



Gephi
makes graphs handy
Download Blog Wiki Forum Support Bug tracker
Home Features Learn Develop Plugins Services Consortium

The Open Graph Viz Platform

Gephi is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free.

Runs on Windows, Mac OS X and Linux.

Learn More on Gephi Platform »

Download FREE
Gephi 0.9.1
Release Notes | System Requirements
▶ Features
▶ Quick start
▶ Screenshots
▶ Videos

Support us! We are non-profit. Help us to innovate and empower the community by donating only 8€:

APPLICATIONS

✓ **Exploratory Data Analysis**: intuition-oriented analysis by networks manipulations in real time.

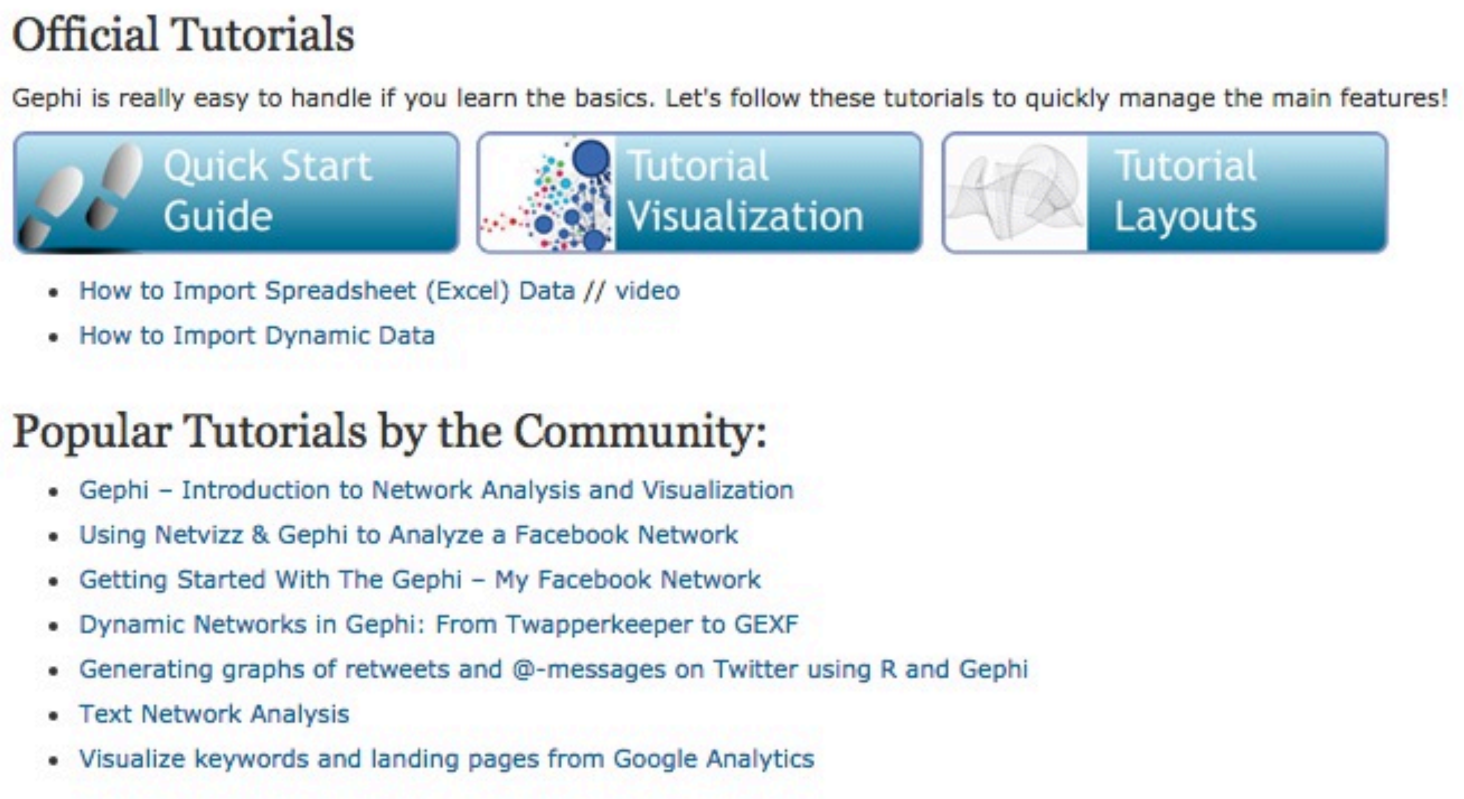
Like Photoshop™ for graphs.

— the Community

PAPERS

Gephi : An Open Source Software for Exploring and Manipulating Networks

Gephi Tutorials



Official Tutorials

Gephi is really easy to handle if you learn the basics. Let's follow these tutorials to quickly manage the main features!

- How to Import Spreadsheet (Excel) Data // video
- How to Import Dynamic Data

Popular Tutorials by the Community:

- Gephi – Introduction to Network Analysis and Visualization
- Using Netvizz & Gephi to Analyze a Facebook Network
- Getting Started With The Gephi – My Facebook Network
- Dynamic Networks in Gephi: From Twapperkeeper to GEXF
- Generating graphs of retweets and @-messages on Twitter using R and Gephi
- Text Network Analysis
- Visualize keywords and landing pages from Google Analytics

http://gephi.org/users