Integrating NetworKit into a web-based Environment for Network Analysis and Exploration

Jörn Kreutel, Beuth University of Applied Sciences Berlin
@NetworKit Day, October 15, 2020
Contents

- Background and Motivation
- Functionality
- Architecture
- Current State and Open Issues
- Demo
Background and Motivation
Domain Background and Motivation

- Research on **bio-bibliographical networks** as a *work in progress*

- **Partners**: Humboldt-University / German Literature Studies, Berlin-Brandenburg Academy of Sciences and Humanities

- Three relatively large bibliographical datasets
  - *Bibliographical Yearbooks*: bibliography of all **fictional literature publications** in SOZ+GDR (1945-1990),
    - appr. 50,000 entries
  - *Bibliography of German Literature and Language Studies and International Bibliography of German Literature and Literary Studies*: international bibliographies of **scientific publications**
    - appr. 320,000 - 350,000 entries

- **Enhancement** of bibliographical data with **biographical information** on authors (e.g. date of birth, affiliation with institutions, places of living etc.), mainly taken from **GND authority files**

- Since October 2019, creation of a dataset of **detailed bibliographical data** on GDR authors (*Research platform „Literary field GDR“*: authors, works, networks, funded by DFG)

- For many aspects of the talk, see Kreutel 2019
Objective, Method, Tools

- Separate subjects of analysis, **common objectives** and analysis methods
  - main focus: **copublication networks** based on contributions to collections and anthologies
  - Maximal network sizes of about 15,000-20,000 nodes and 300,000-600,000 edges, based on 3000-10,000 collections as network constituting entities

- **Objective**: Insight into a discipline‘s / a field‘s historical evolution, main agents and groups of agents, publication types, relevance of institutions, patterns of careers, contributions of private/professional networks etc. (following, e.g., Bourdieu 1999; De Nooy 2003; Bottero and Crossley 2011)

- **Method**: Social Network Analysis (SNA) + “conventional” Statistics

- What **tools** do we need for investigation?
Starting Point: NetworKit

- NetworKit had been found as a potential core engine for SNA in 2015
- Installation of NetworKit on MacOS in Summer 2016

“Now the SNA core engine is running, what else do we need in order to do SNA on a particular domain using some domain specific data set?”
- How can the graphs to be analysed by NetworKit be created?
- How can the graphs and the analysis results be visualised?
- How can the visualisation be explored and enriched by domain specific data to qualitatively augment the quantitative SNA results?

Motivation of enrichment:
- Graphs for SNA need to be extracted from raw data using appropriate queries
- But: loss of detail information by merging of multiple aspects of raw data into the single dimension of a relation between agents
Raw Data vs. Graph Data

Graph itself only represents information on related entities and those entities that constitute relations ("edge values").
Graph Data Enrichment

How can we seamlessly mediate between a “distant reading” (Moretti 2005) perspective of analysis results (e.g. based on entities’ KPIs calculated by SNA tools, like centrality) to a “close(r) reading” of an entity’s details (e.g. their attributes beyond the given network)?
Issues to be resolved

1. How can creation, analysis and visualisation of graphs be integrated into a seamless workflow?

2. How can information represented by a network graph be enriched by linkable data?

3. How can domain dependent enrichment functionality be supported by the generic implementation of the graph visualisation component?
Approach

- Necessity of a **generic software solution** as an alternative to “scripting-on-demand”
  
  - Fast **portability** to new domains and datasets
  
  - **reusability** of analyses for different datasets with identical structure

- Overall **objective**: provision of functionality via a **web-based service and UI** without necessity of on-site installation

- Related technical work:
  - Gephi and R support creation and analysis of networks based on SPARQL queries (Totet 2015; van Hage)
  - Gephi allows additional **custom attributes** for nodes and edges and foresees UI extension points for **custom java plugins**
  - BUT: **No** “rich” interactive **web based UI** provided
Functionality
Development History

- Starting in late summer 2016: “Proof of Integration” for SPARQL+SNA+Web-based visualisation

- From the beginning: usage of NetworKit as SNA tool (Staudt et al. 2015)

  “Building a web-based Environment for Network Analysis and Exploration on top of NetworKit”

- Initial Scenario: Co-Occurrence networks of drama characters based on “Projekt Gutenberg” data (see, e.g., Trilcke et al.)

- Application to the above mentioned bio bibliographical data sets

- Since then continuous enhancement due to proven practicability and expressivity of analyses in the area of bio-bibliographical data (see Kreutel et al. 2019)
Key Features

1. Integrated **Workflow and UI**
   - **UI-based configuration** of query and analysis details (query, data sources, analysis parameters)
   - integrated **processing pipeline** for SPARQL etc. querying, graph creation, SNA and visualisation, **triggered by a single action** at UI level

2. Domain specific **Enrichment** of graph data and visualisation
   - **Network Manipulation**: manipulation of graph visualisation based on domain specific KPIs for the entities represented by the graph’s nodes (e.g. node size based on number of overall publications)
   - **Network Browsing**: exploration of nodes and edges by providing domain specific views of entities based on linkable data
   - **Network Statistics**: applying analysis templates to selected nodes and/or edges of the graph (e.g. for obtaining insight into qualitative features of detected communities as a potential source of community formation); correlating domain specific KPIs with SNA KPIs

3. Enrichment as **Scripting**
   - Enrichment functionality is implemented in **JavaScript** and managed via the UI
Enrichment Example

Network Analysis, Jörn Kreutel @NetworKit Day 2020
Effect of Enrichment

- **re-integration of data** that has been abstracted over by graph creation

- enrichment may cover **arbitrary domain specific data** that can be linked with the entities contained within the graph

- **separation of concerns** between graph data and analysis and enrichment functionality (vs. custom attributes on the graph itself)
Architecture
System Architecture and Implementation

Unless otherwise stated, inter-component communication uses HTTP.
A Web-based Platform for SNA?

1. Configuration, Workflow, Data Connectors
   - Node.js Server
     - Graph Creation and Analysis Workflow
     - Extended Analysis Workflow
     - Configuration Storage
   - MongoDB Server
     - Configuration Data
   - Apache Tomcat Server
     - Query Service
     - GraphData Access Service
   - Shared Drive
     - Graph Data

2. Domain Data
   - GraphDB Server
     - RDF Repository
   - Data Access Service N.N.
     - Dataset N.N.

3. Network Analysis
   - Python HTTP Server
     - NetworkKit
   - Browser
     - User Interface
     - Graph Visualisation

Domain Experts, Data Engineers

enrichment (option)
Current State and Open Issues
Current State and Open Issues

- **Domain independent solution** applied to and further developed on the basis of two example domains and various datasets
  - Straight ad-hoc usage for **analysing new domains** and datasets (e.g. Wikidata movies and actors)

- **BUT:** So far, **no self service by domain experts** for graph exploration and analysis (additionally backed by DH engineers doing setup and developing extension scripts).

- Both UI and underlying implementation would benefit from some cleanup.

- Large Graphs seriously challenge **performance of browser-based visualisation** both on the basis of SVG (D3) and WebGL (sigma.js).

- Current environment is rather a **prototype for a web based SNA platform** providing integration and visualisation services for SNA on the basis of NetworKit than such a platform itself.
Thank you!

joern.kreutel@beuth-hochschule.de
References

- San Martin, M., Gutierrez, C.: Representing, Querying and Transforming Social Networks with RDF/SPARQL. In: European Semantic Web Conference. Springer (2009), pp. 293–307