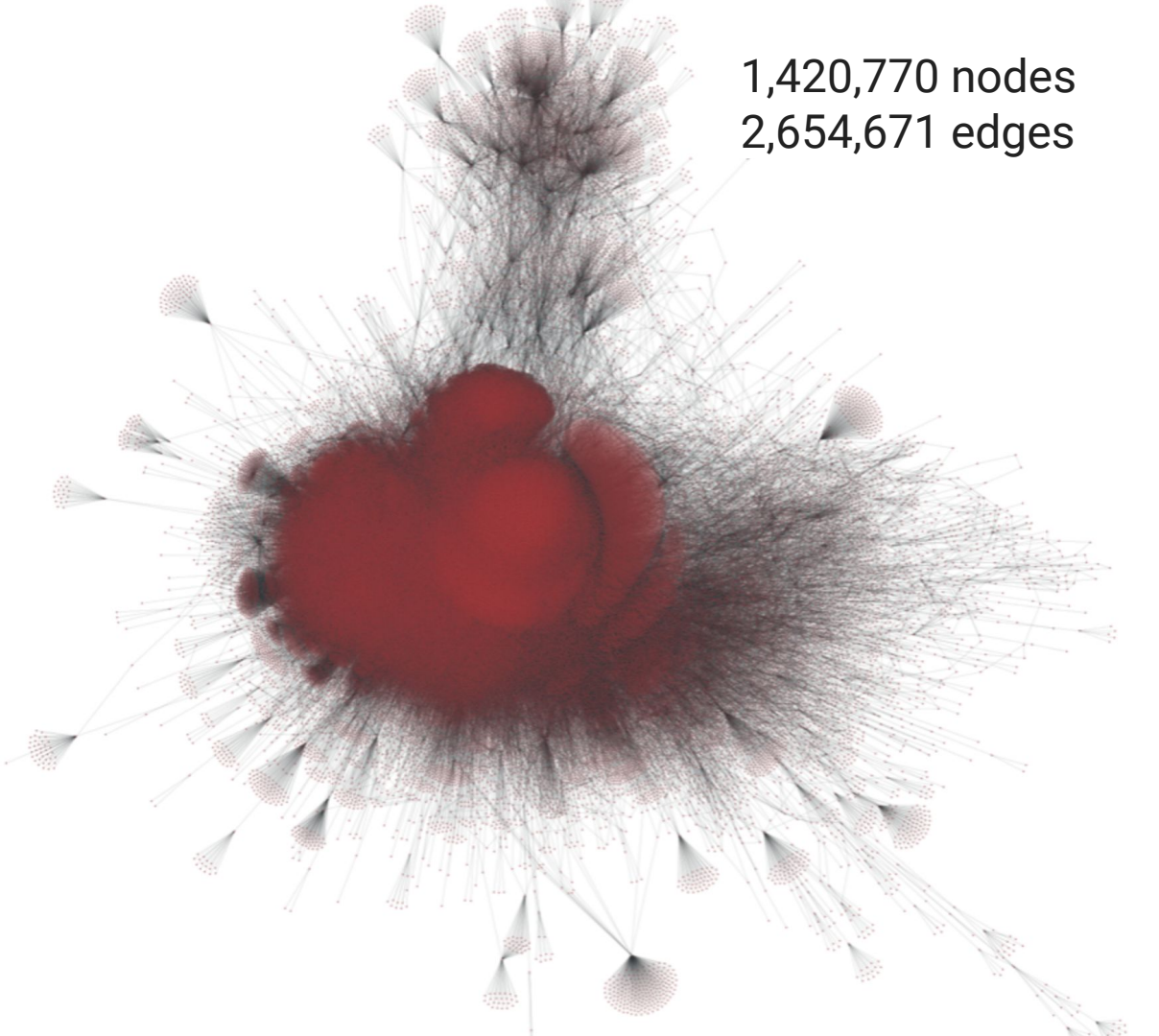
Network analysis of large Knowledge Graphs

Daniel Garijo
Universidad Politécnica de Madrid
daniel.garijo@upm.es
@dgarijov
(with some slides from Pedro Szekely)

Starting from a KG

What kinds of analysis can we do with it?

- Most relevant nodes?
- Paths between nodes?
- Communities?
- Property subgraphs?



Typical network analysis operations: node centrality and pagerank

Node centrality counts the in/out degree of each node in the graph

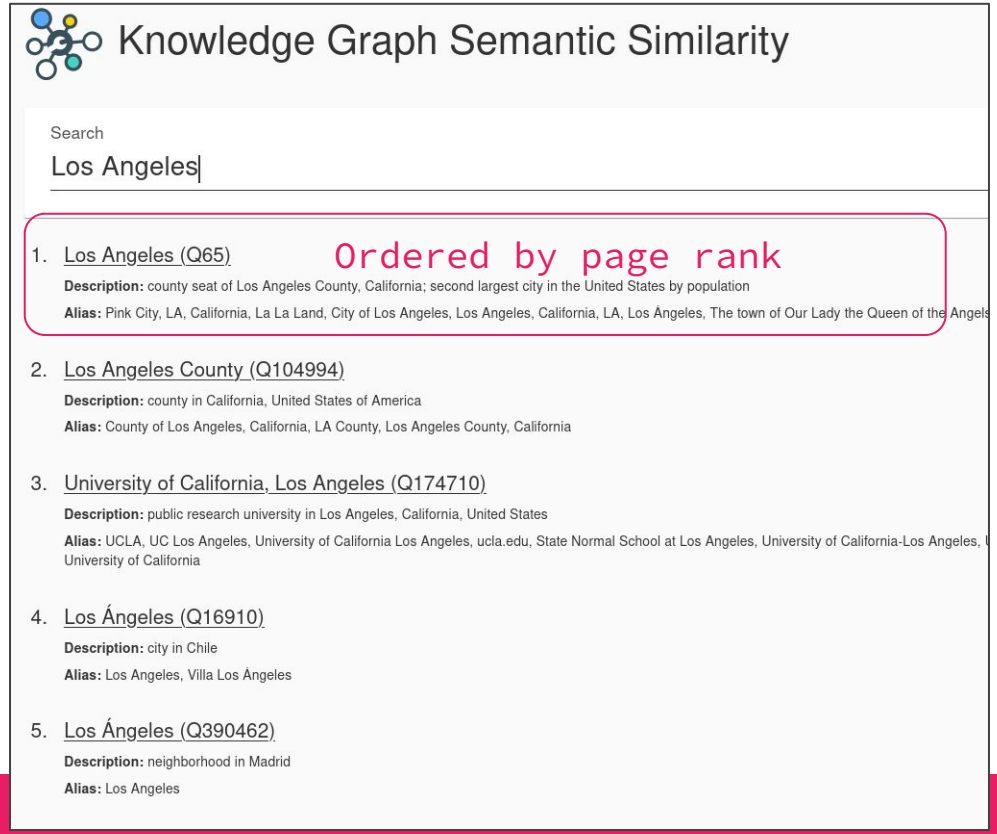
Page rank assigns a weight to each node based on the links pointing to that node. Higher page rank -> higher relevance

https://en.wikipedia.org/wiki/File:PageRank-hi-res.png

Typical network analysis operations: node centrality and pagerank

Why is page rank useful?

- Returning **relevant nodes** in case of ambiguous search terms
- Highlight most “relevant” entities in the graph

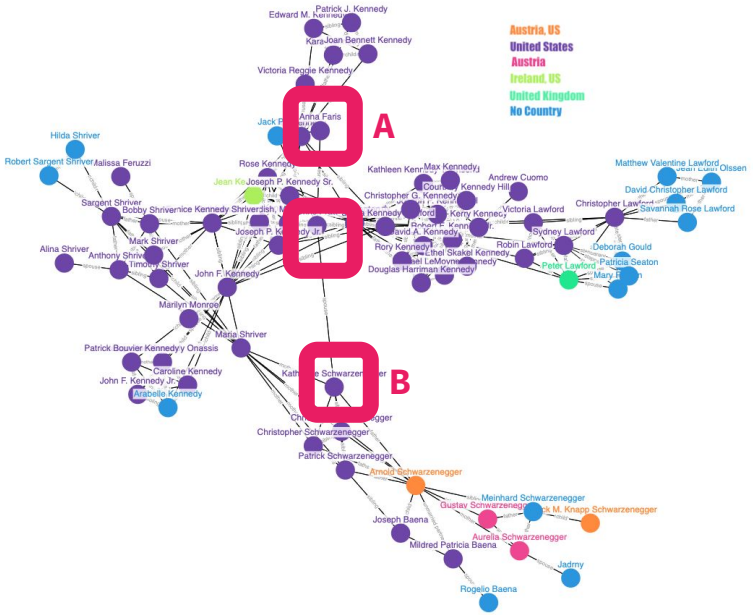


Typical network analysis operations: shortest paths

Given a set of nodes, what is the **shortest path** between them in the graph?

Why find shortest paths?

- Find **a connection** between two entities in the graph
- Highlight hidden commonalities between nodes

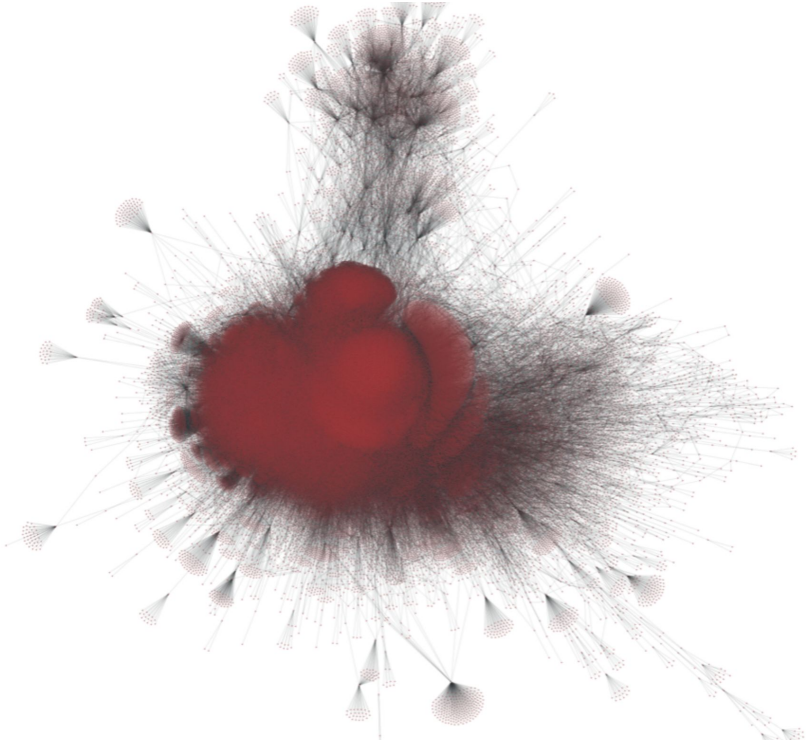


Typical network analysis operations: Connected Components

A component is set of nodes which **are connected** by one or multiple edges

Why is it useful?

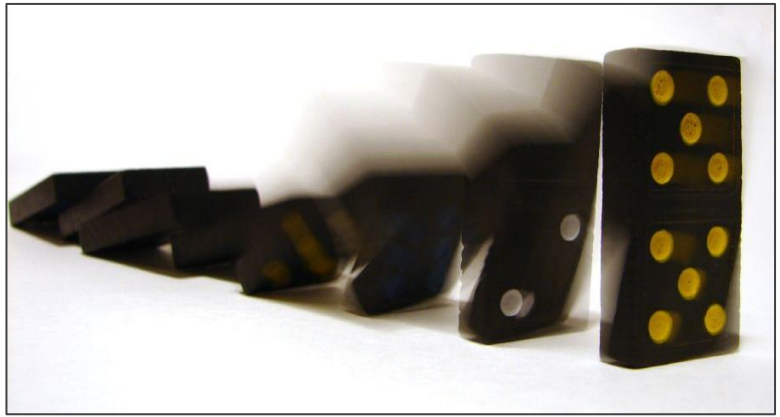
- To know more about the distribution of the graph
- Starting point to detecting highly connected communities
- Highlight potential missing information linking disconnected components
- Locate disconnected nodes



Typical network analysis operations: Reachable Nodes

Which nodes can be reached given:

- A set of **root nodes**
- Using a fixed **set of properties**



Why is it useful?

- Creating subsets of the graph of interest
  - families (spouses, children, parents, etc.)
  - political parties
  - supervisor-PhD student networks
  - etc.

Network analysis

Expensive! You can’t use SPARQL to run these queries in large KGs

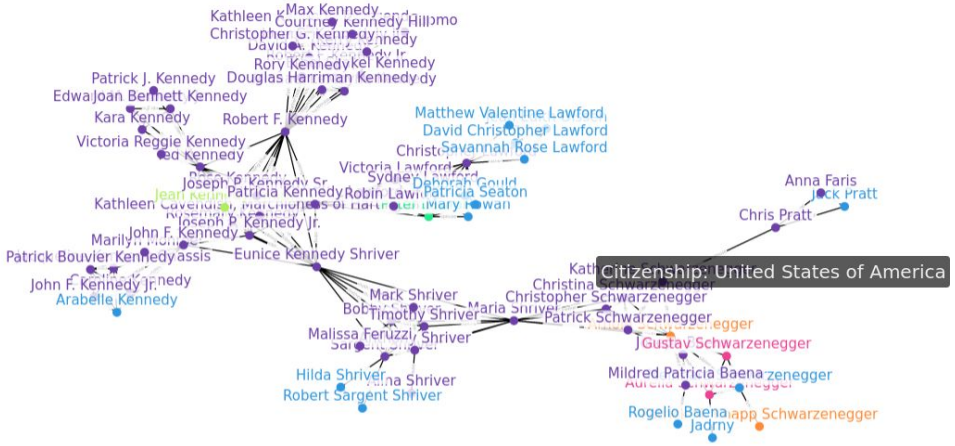
- Some queries have to explore large portions of the graph

Visualization of results is usually performed with **separate libraries**, in different formats

Visualization is key!

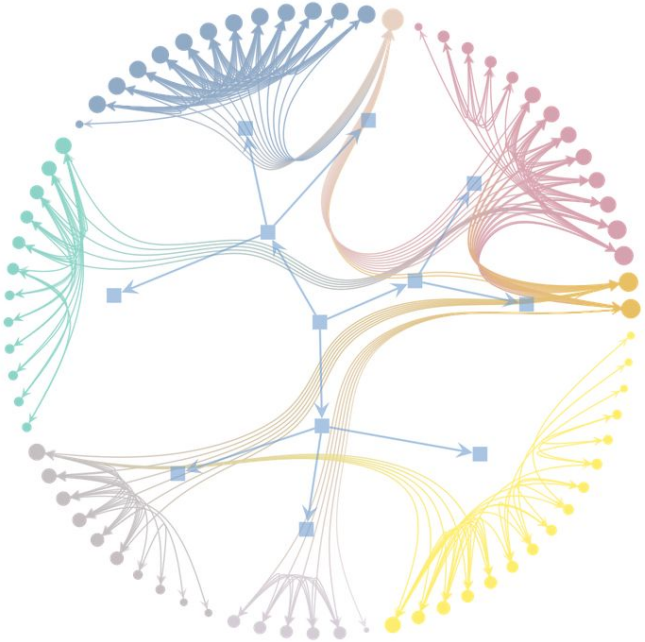
Critical to understand the different relationships available in the obtained results.

Below are some visualizations from the same graph:

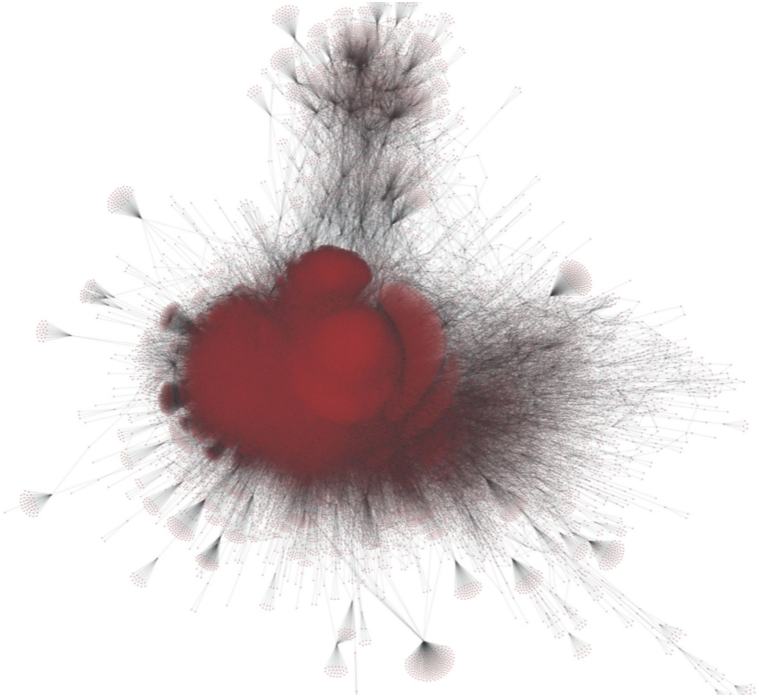


Visualization is key!

Hierarchical representation

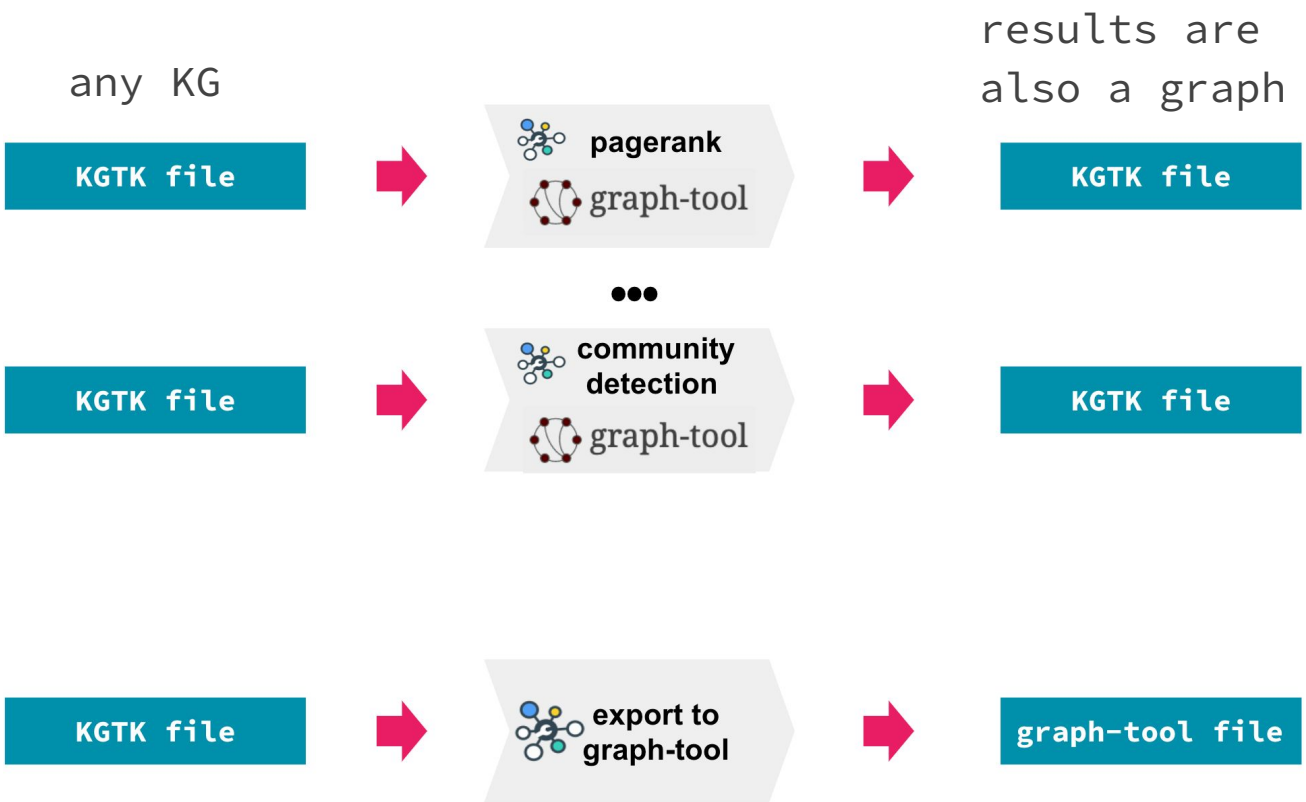


Connected components



KGTK wraps graph-tool for network analysis

https://graph-tool.skewed.de/

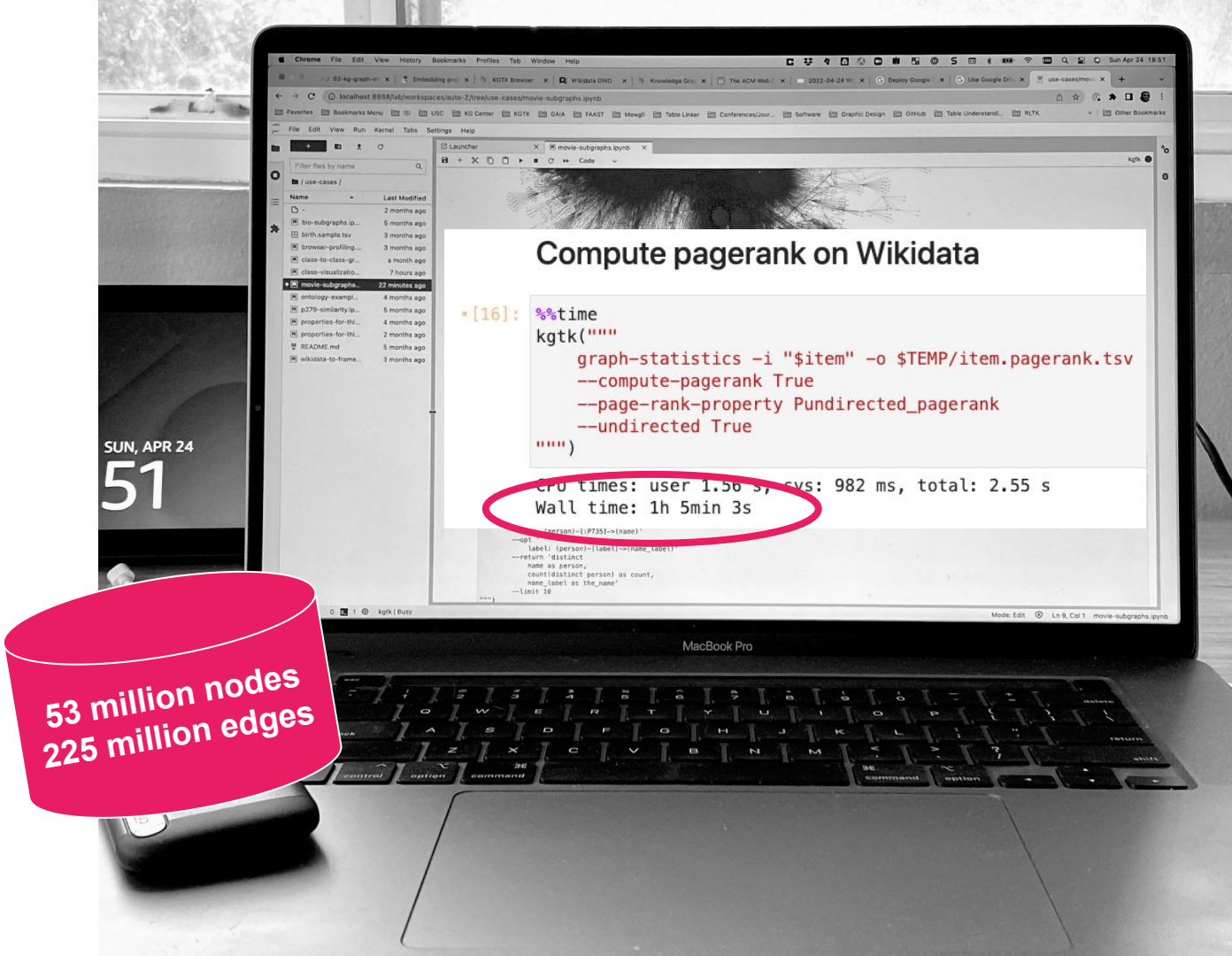


convenient

efficient

maximum flexibility

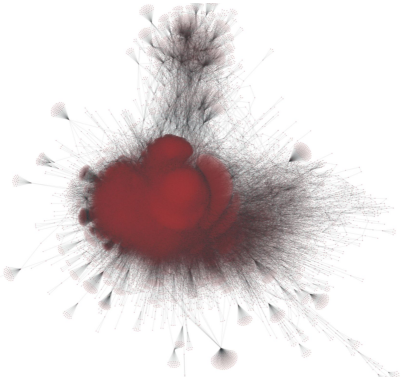
KGTK network analysis is efficient and scalable



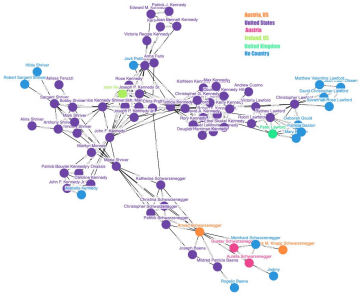
KGTK wraps graph-tool vasturiano / force-graph for network visualization

any KG

KGTK file → export to graph-tool → large-scale visualization graph-tool →



KGTK file → interactive visualization / force-graph →



KGTK uses the best tools for the job

KGTK vs Neo4J approach to network analytics

KGTK	Neo4J
Wraps existing libraries (https://graph-tool.skewed.de/)	Re-implements algorithms (https://github.com/neo4j-contrib/neo4j-graph-algorithms)
Efficient C/C++ implementation, Python API	Java
Comprehensive coverage of algorithms e.g., flow and spectral algorithms available	Rich, but limited set of algorithms
Open architecture	Proprietary architecture
Ability to add new algorithms easily	Difficult to add new algorithms
Pipeline integration	Tight integration

Let’s jump to the notebook

Folder with **all notebooks**:
https://github.com/usc-isi-i2/kgtk-notebooks

Network analysis notebook:
https://colab.research.google.com/drive/1Lat732XpHv1RMswoYsz_wUk_6eKEf8Xt?usp=sharing (remember to save it in your Gdrive)

Note: Graph tool will not work in colab!

KGTK integrates state of the art tools for network analysis and visualization

KGTK file

Full graph
Subset (after filtering)

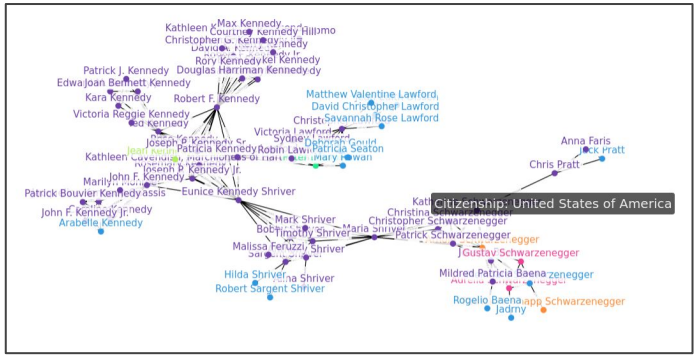
KGTK

visualization

Html file (ktkg-visualize)
Graph Tool file (export-gt)

	node1	label	node2	node1;label	label;label	node2;label
0	Q1086823	P22	Q345517	'Christopher Lawford'@en	'father'@en	'Peter Lawford'@en
1	Q1086823	P25	Q432694	'Christopher Lawford'@en	'mother'@en	'Patricia Kennedy Lawford'@en
2	Q1086823	P26	Q75326809	'Christopher Lawford'@en	'spouse'@en	'Jean Edith Olssen'@en
3	Q1086823	P3373	Q75326777	'Christopher Lawford'@en	'sibling'@en	'Victoria Lawford'@en
4	Q1086823	P3373	Q75326779	'Christopher Lawford'@en	'sibling'@en	'Sydney Lawford'@en
...	...	...	...	...	...	...
489	Q9696	P40	Q230303	'John F. Kennedy'@en	'child'@en	'Caroline Kennedy'@en
490	Q9696	P40	Q316064	'John F. Kennedy'@en	'child'@en	'John F. Kennedy Jr.'@en
491	Q9696	P40	Q3290402	'John F. Kennedy'@en	'child'@en	'Patrick Bouvier Kennedy'@en
492	Q9696	P40	Q75326753	'John F. Kennedy'@en	'child'@en	'Arabelle Kennedy'@en
493	Q9696	P451	Q4616	'John F. Kennedy'@en	'unmarried partner'@en	'Marilyn Monroe'@en

494 rows × 6 columns



KGTK wraps graph-tool for network analysis

https://graph-tool.skewed.de/