

# Creating subsets of Wikidata

Pedro Szekely  
University of Southern  
California

# Creating subgraphs of Wikidata is hard

Hard to specify what I want in the subgraph

- Do I specify what I want, what I don't want, or both?
- I don't want to say too much because I will get it wrong

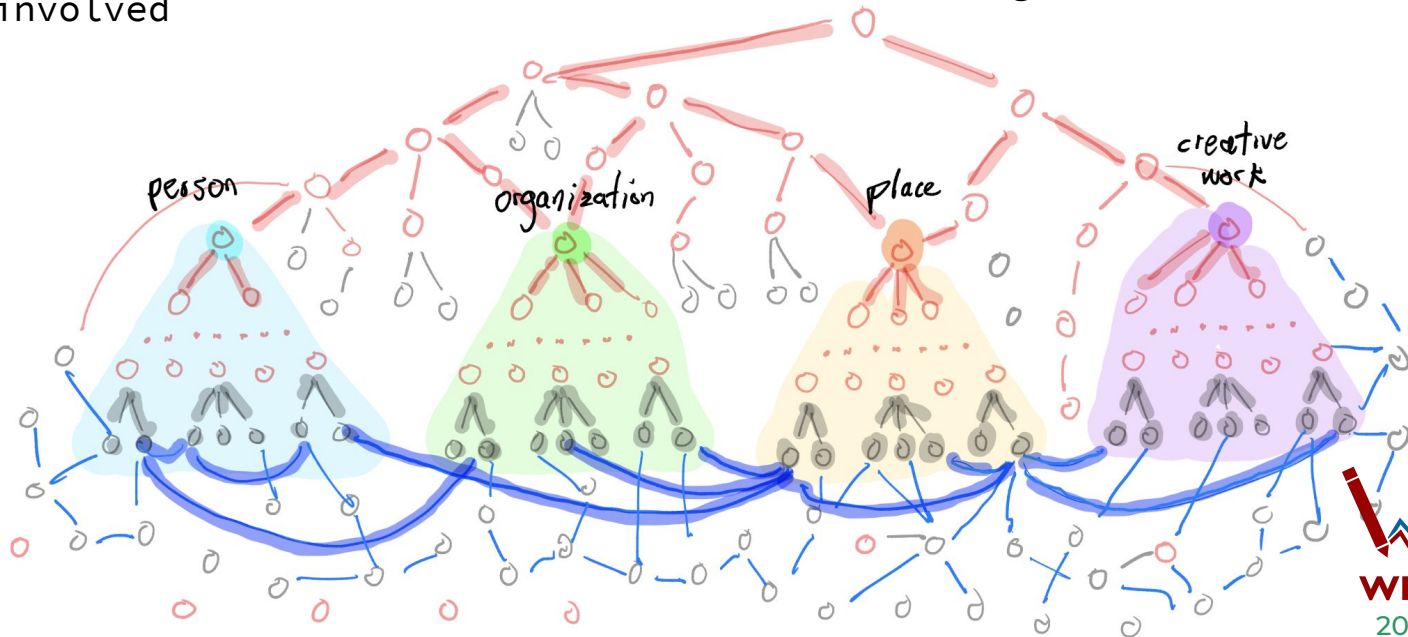
I want to get a coherent subgraph

- classes and super classes (P31, P279)
- property definitions
- qualifiers and references
- labels, aliases and descriptions for all items in my subgraph

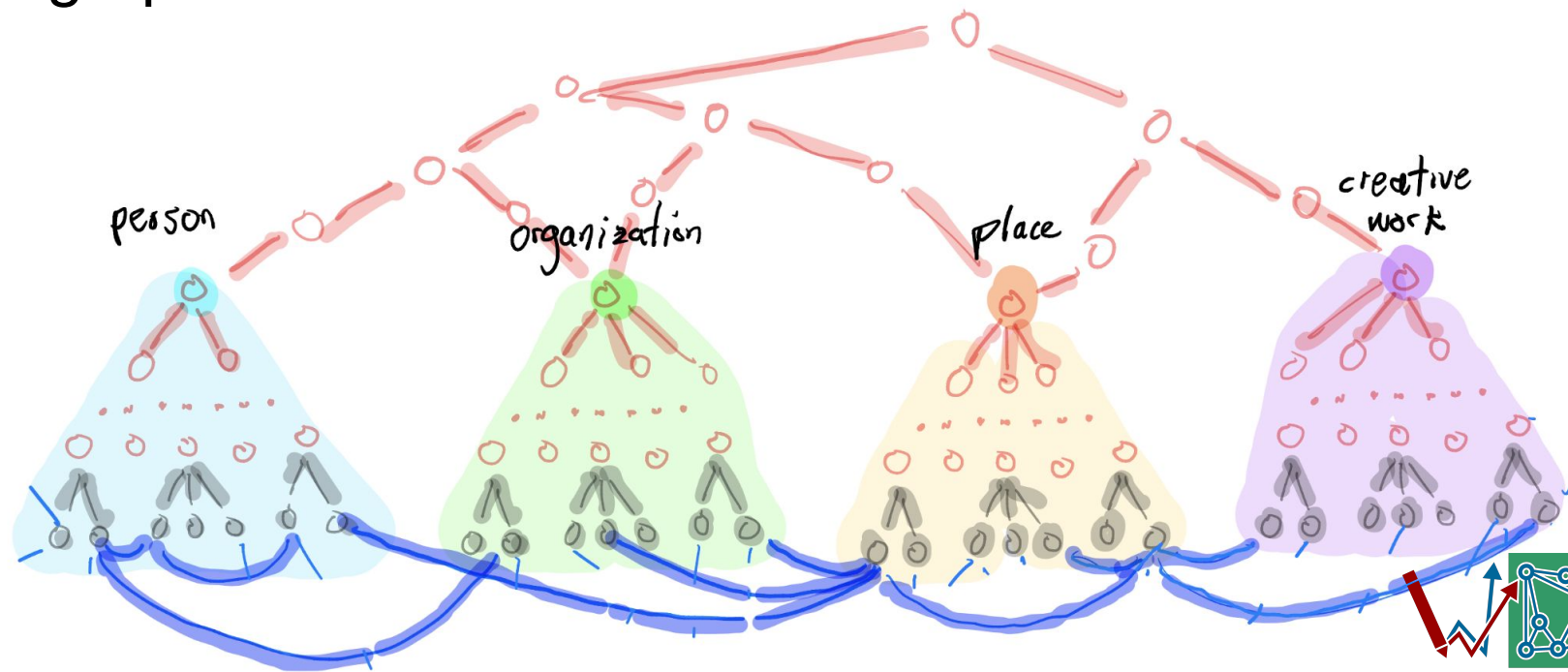
# I want a graph with about movies and books

I want info about the people, organizations and places involved

include "subclass of" edges to root



# Wikidata person/organization/place/creative work subgraph

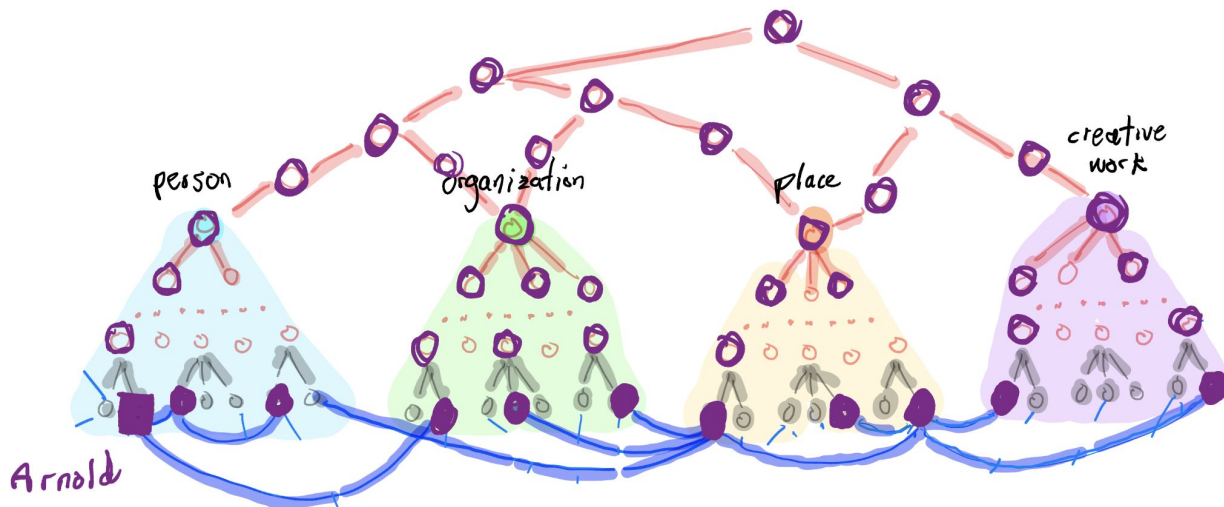


4 hours on my laptop

# Now I want an Arnold Schwarzenegger graph

- Start from Arnold
- Go forward K (e.g. 3) hops
- Get everything connected from there

| what               | count     |
|--------------------|-----------|
| edges (all)        | 2,614,950 |
| edges (qualifiers) | 443,899   |
| items              | 58,522    |
| properties         | 3,831     |
| classes            | 14,490    |



<https://github.com/usc-isi-i2/kgtk/blob/master/tutorial/build-kg/build-tutorial-graph.ipynb>



# Arnold Schwarzenegger

(Q2685)

Arnie | Arnold Alois Schwarzenegger | Arnold Strong | Governor | Governor Schwarzenegger | Schwarzenegger | The Styrian Oak | The Terminator  
 Austrian-American actor, businessman, bodybuilder and politician

## Properties

|                         |  |
|-------------------------|--|
| instance of             | human  |
| sex or gender           | male   |
| birth name              | Arnold Alois Schwarzenegger [German]   |
| name in native language | Arnold Alois Schwarzenegger [German]   |
| nickname                | Arnie [mul]<br>Governor [English]<br><i>start time</i> 2003<br>The Austrian Oak [English]<br>The Styrian Oak [English]<br><i>literal translation</i> Roble Austriaco [Spanish] |
| pseudonym               | Arnold Strong   ``governator``   |
| date of birth           | July 30, 1947  |
| work period (start)     | 1969   |
| child                   | Christina Schwarzenegger<br>Christopher Schwarzenegger<br>Joseph Baena<br>Katherine Schwarzenegger<br>Patrick Schwarzenegger   |
| sibling                 | Meinhard Schwarzenegger  |
| spouse                  | Maria Shriver<br><i>start time</i> April 26, 1986<br><i>end time</i> July 01, 2011   |
| relative                | Patrick M. Knapp Schwarzenegger  |

## Gallery

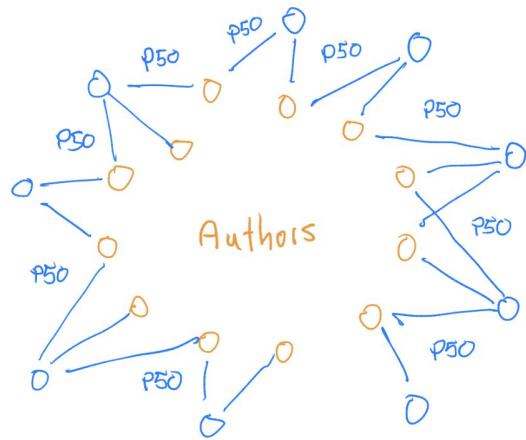
## Identifiers

|                                     |                       |
|-------------------------------------|-----------------------|
| abART person ID                     | 79216                 |
| Acharts.co artist ID                | arnold_schwarzenegger |
| Alexander Turnbull Library ID       | 118183                |
| Alcinema person ID                  | 11017                 |
| AllMovie person ID                  | p110501               |
| AlloCiné person ID                  | 1067                  |
| Amazon author ID                    | B000AP7VZW            |
| Amazon Music artist ID              | B0013691AA            |
| American Film Institute person ID   | 223815                |
| Babelio author ID                   | 237798                |
| Ballotpedia ID                      | Arnold_Schwarzenegger |
| Behind The Voice Actors person ID   | Arnold-Schwarzenegger |
| BFI Filmography person ID           | 187025                |
| BFI Films, TV and people ID         | 4ce2b9fbd2853         |
| Biblioteca Nacional de España ID    | XX974740              |
| Bibliothèque nationale de France ID | 139419911             |
| BIBSYS ID                           | 90851780              |
| Box Office Mojo person ID           | arnoldschwarzenegger  |
| Brockhaus Enzyklopädie online ID    | schwarzenegger-arnold |
| C-SPAN person ID                    | arnoldschwarzenegger  |
| CANTIC ID                           | a10891948             |
| CCAB ID                             | 000268391             |
| Cinema.de ID                        | 1567829               |
| CineMagia person ID                 | 459                   |
| cinematografo.it name or company ID | 110845                |
| Cineuropa person ID                 | 261714                |
| CINii author ID (books)             | DA1033823X            |



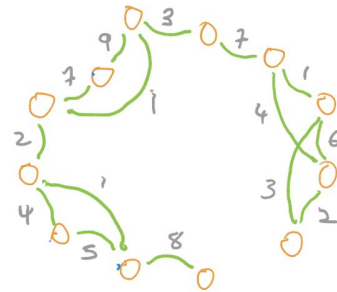
# I want to create a coauthor subgraph

Papers (P50:author)



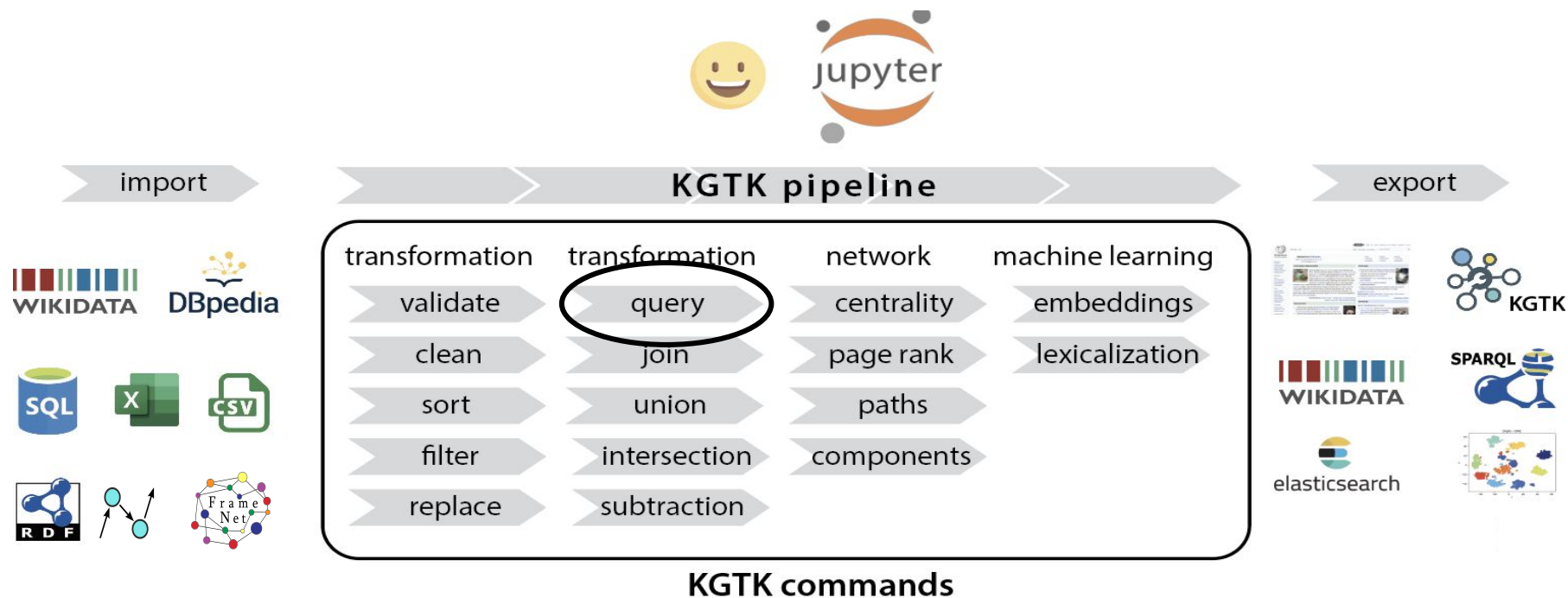
Coauthor Graph

with counts of coauthored papers



The new subgraph uses a **coauthor** property that is not present in Wikidata

# KGTK - Knowledge Graph Toolkit: Rich Support For Working With Any KG



<https://github.com/usc-isi-i2/kgtk>



## KGTK runs faster on my laptop than SPARQL on a server

| Query            | Kypher 16GB laptop | Kypher 32GB laptop | SPARQL 256GB local server | SPARQL public |
|------------------|--------------------|--------------------|---------------------------|---------------|
| First names      | 24.37              | 8.28               | 31.05                     | time out      |
| Class instances  | 104.97             | 88.97              | >24 hours                 | time out      |
| Film instances   | 0.03               | 0.04               | 1.91                      | time out      |
| Author network   | 61.55              | 66.39              | >24 hours                 | time out      |
| Cancer network   | 3.18               | 2.62               | 40.19                     | time out      |
| ULAN identifiers | 0.56               | 0.20               | 1.08                      | *             |
| DBpedia spouses  | 3.92               | 3.43               | n/a                       | n/a           |

memory is RAM, all times in minutes except noted otherwise, (\*) error, query too large

# I can create a coauthor subgraph in 1 hour on my laptop

kgtk query -i wikidata  
--match

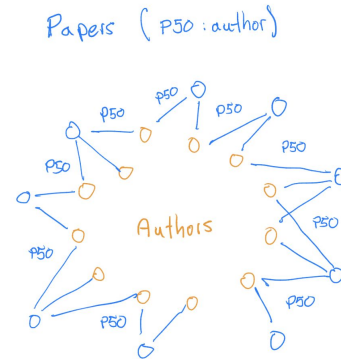
```
(pub) - [:P31] -> (class),  
(class) - [:P279star] -> (:Q591041),  
# Q591041: scientific publication  
(pub) - [:P50] -> (author1),  
(pub) - [:P50] -> (author2)  
# P50: author
```

--where author1 > author2

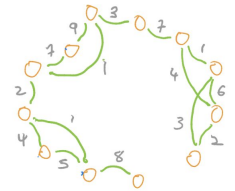
--return

```
distinct author1 as node1,  
"Pcoauthor" as label,  
author2 as node2,  
count(distinct pub) as count_publications
```

| node1     | label     | node2     | count | node1;label    | node2;label        |
|-----------|-----------|-----------|-------|----------------|--------------------|
| Q60320900 | Pcoauthor | Q60394812 | 396   | 'Jorge ...'@en | 'Hagop Kan ...'@en |
| Q60394812 | Pcoauthor | Q66370727 | 236   | 'Hagop ...'@en | 'Susan O'Brien'@en |
| Q40614280 | Pcoauthor | Q60394812 | 186   | 'Farha ...'@en | 'Hagop Kan ...'@en |



Coauthor Graph  
with counts of coauthored papers



# Summary

Create sophisticated subgraphs using KGTK

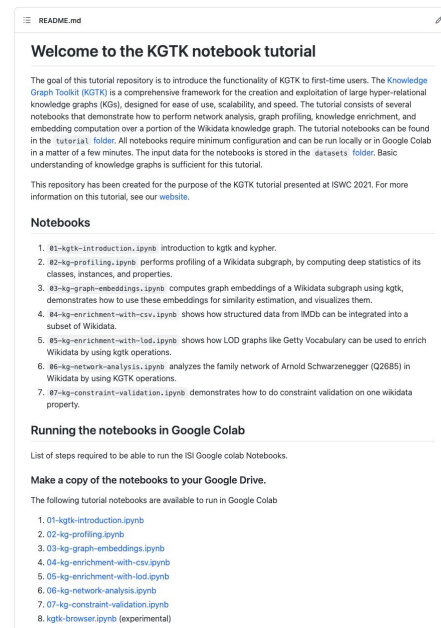
KGTK is faster on a laptop than SPARQL on a server

KGTK can do much more than subgraphs

- query
- knowledge graph profiling
- network analytics
- embeddings

<https://github.com/usc-isi-i2/kgtk>

Feb 15, 2021 snapshot of Wikidata (minus scientific publications): [Google Drive](#)



The screenshot shows the README.md file for the KGTK notebook tutorial. It includes a welcome message, a description of the tutorial's goal, a list of notebooks, and instructions on how to run them in Google Colab.

**Welcome to the KGTK notebook tutorial**

The goal of this tutorial repository is to introduce the functionality of KGTK to first-time users. The Knowledge Graph Toolkit (KGTK) is a comprehensive framework for the creation and exploitation of large hyper-relational knowledge graphs (KGs), designed for ease of use, scalability, and speed. The tutorial consists of several notebooks that demonstrate how to perform network analysis, graph profiling, knowledge enrichment, and embedding computation over a portion of the Wikidata knowledge graph. The tutorial notebooks can be found in the `tutorials/` folder. All notebooks require minimum configuration and can be run locally or in Google Colab in a matter of a few minutes. The input data for the notebooks is stored in the `datasets/` folder. Basic understanding of knowledge graphs is sufficient for this tutorial.

This repository has been created for the purpose of the KGTK tutorial presented at ISWC 2021. For more information on this tutorial, see our [website](#).

### Notebooks

- 01-kgtk-introduction.ipynb: introduction to kgtk and kgkter.
- 02-kg-profiling.ipynb: performs profiling of a Wikidata subgraph, by computing deep statistics of its classes, instances, and properties.
- 03-kg-graph-embeddings.ipynb: computes graph embeddings of a Wikidata subgraph using kgk, demonstrates how to use these embeddings for similarity estimation, and visualizes them.
- 04-kg-enrichment-with-csv.ipynb: shows how structured data from MDBs can be integrated into a subset of Wikidata.
- 05-kg-enrichment-with-lod.ipynb: shows how LOD graphs like Getty Vocabulary can be used to enrich Wikidata by using kgk operations.
- 06-kg-network-analysis.ipynb: analyzes the family network of Arnold Schwarzenegger (Q2685) in Wikidata by using KGTK operations.
- 07-kg-constraint-validation.ipynb: demonstrates how to do constraint validation on one wikidata property.

### Running the notebooks in Google Colab

List of steps required to be able to run the ISI Google colab Notebooks.

**Make a copy of the notebooks to your Google Drive.**

The following tutorial notebooks are available to run in Google Colab

- 01-kgtk-introduction.ipynb
- 02-kg-profiling.ipynb
- 03-kg-graph-embeddings.ipynb
- 04-kg-enrichment-with-csv.ipynb
- 05-kg-enrichment-with-lod.ipynb
- 06-kg-network-analysis.ipynb
- 07-kg-constraint-validation.ipynb
- kgtk-browser.ipynb (experimental)

[Try KGTK on Google Colab](#)

colab

# Thanks for your attention!

Get in touch with us:

**Pedro Szekely**

[szekely@usc.edu](mailto:szekely@usc.edu)

<https://usc-isi-i2.github.io/szekely/>

**Credits**

<https://github.com/usc-isi-i2/kgtk>



**WIKI DATA CON**

2021 - a sustainable  
future for Wikidata