Centrality Measures and Knowledge Graphs

CS 579 Online Social Network Analysis

Dr. Cindy Hood
9/18/25

Homework Assignments

- HW #2 assigned Due by midnight Friday 9/19
- HW #3 assigned later this week
  - You may work in groups up to 4 students (no exceptions) on this hw
- Please contact TAs with questions on hw grading

Exams and Final Project Poster Presentation

- Exam 1 – Oct 9 in class
- Exam 2 – Dec 2 in class

- Final Project Poster Session – Dec 4 in class

- Online students (sections 2 and 3) will have remote options

Teaching Assistants

- Siva Krishna Golla
  - sgolla2@hawk.illinoistech.edu
  - Mondays 2-3pm on zoom
- Khush Dhiren Patel
  - kpatel210@hawk.illinoistech.edu
  - Wednesdays 11-12 online
- Aswith Sama
  - asama@hawk.illinoistech.edu
  - Thursdays 3-4pm on zoom

- **Not yet officially working, waiting for authorization (US govt)**

HW #2 due by midnight 9/19

- In this assignment you will create networks/graph models from 2 different datasets. You may use any tool/platform/language that you like. I have attached a few pages from the Elements of Network Science Book that illustrate basic use of Stata, R and Python. Section 2.3 ElementsofNetworkScience.pdf Download Section 2.3 ElementsofNetworkScience.pdf
- (1) The first dataset is Chicago Community Areas https://en.wikipedia.org/wiki/Community_areas_in_Chicago
- Links to an external site.
- Nodes = Community areas
- Edges = Shared physical boundary (i.e. adjacency) with other community area. Note that you may have to make some assumptions here since you are determining boundaries from the image of the map on the page cited above. State your assumptions.
- You will then create a labelled visualization of this graph and plot the degree distribution of the nodes. You will submit
- (1a) Input file with graph representation,
- (1b) Labelled visualization of network created.
- (1c) Plot of degree distribution.

HW #2 con't

- (2) The second dataset is the CS 579 Class Participant Data Social Network Data collection.xlsx
- Links to an external site.
- Nodes = Class Participants, entities in common
- Edges = Shared entity
- You will create a bipartite graph. Some data cleaning will be necessary. State and justify any assumptions you make during the data cleaning. You will then create a unimodal graph that is a projection of the bipartite graph.
- You will create labelled visualizations of both the bipartite and unimodal graphs and plot the degree distribution of the unimodal graph. You will submit
- (2a) Input file for the bipartite graph.
- (2b) Labelled visualization of bipartite graph.
- (2c) Description of method for projecting bipartite graph to unimodal graph including code.
- (2d) Labelled visualization of unimodal graph.
- (2e) Plot of degree distribution of unimodal graph.

HW #2 – con't

- (3) Compare the degree distributions of the graphs from the two different datasets. What is similar? What is different? Is this what you expected? Why or why not?
- (4) Provide the details of how you did this assignment. What tools did you use to complete the assignment? Why did you choose the tool? Provide citations and links to references and code used. If AI (e.g. ChatGPT, etc.) was used, please include a transcript of the exchange.
- The above can be submitted in a zipped folder that includes
- input files labelled as Input_file1, Input_file2
- pdf report of everything else


Chicago
Central
Far North Side
Far Southeast Side
Far Southwest Side
North Side
Northwest Side
South Side
Southwest Side
West Side
A. Burnside
B. Oakland
C. Montclare
O'Hare
Edison Park
Norwood Park
Jefferson Park
Forest Glen
North Park
Rogers Park
West Ridge
Edge-water
Lincoln Square
Albany Park
Uptown
Portage Park
Irving Park
Dunning
North Center
Lakeview
Avondale
(C)
Belmont Cragin
Hermosa
Logan Square
Lincoln Park
Austin
Humboldt Park
West Town
Near North
West Garf. Park
East Garfield Park
Near West Side
Loop
North Lawndale
Near South
Lower West Side
South Lawndale
Bridge-Port
Armor Square
Douglas
McKinley Park
(B)
Archer Heights
Brigton Park
New City
Fuller Park
Grand Blvd
Ken-Wood
Garfield Ridge
West Elsdon
Gage Park
Wash-ington Park
Hyde Park
Clearing
West Lawn
Chicago Lawn
West Engle-Wood
Engle-Wood
Woodlawn
Greater Grand Crossing
South Shore
Ashburn
Auburn Gresham
Chatham
Avalon Park
South Chicago
(A)
Calumet Heights
Wash-ington Heights
Beverly
Roseland
Pullman
East Side
South Deering
Mount Greenwood
Morgan Park
West Pullman
Riverdale
Hegewisch

Why are we modeling information in a network/graph structure?

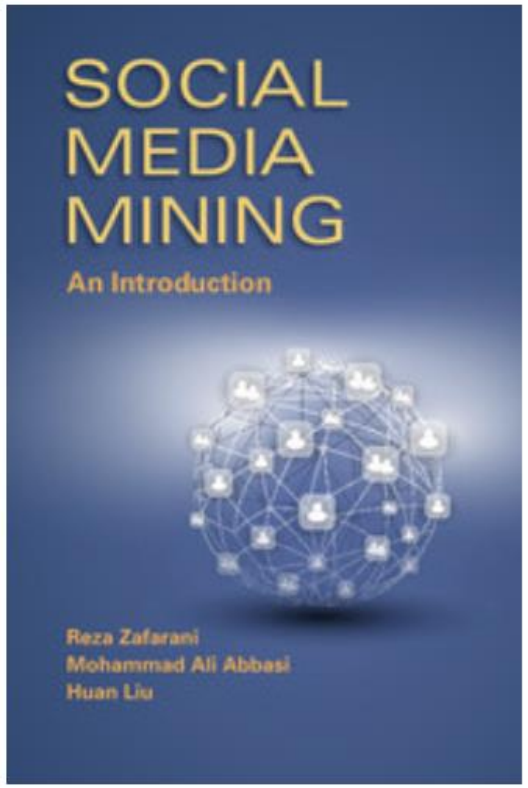
- HW#2
- To see if we can use the network/graph to help
  - Identify relationships
  - View structures
  - Identify patterns
  - Etc

What information helps (i.e. should be used in the model if possible)?

- Information that distinguishes members of the population from one another and characterizes more than 1 member of the population

Languages spoken (separate with comma)
English, Gujarati, Hindi
English, Gujarati, Hindi
English, Mandarin, Cantonese
English, French, Spanish, Chinese
English, Bengali,Hindi,Kannada
English, Hindi, Gujarati
English, Tamil, Tagalog
English
English, Chinese, Japanese
English, Chinese, Japanese
English
English, Urdu, Arabic, Pashto, Mandarin
English,Arabic
Indonesian, English, Mandarin

References



http://www.socialmediamining.info

Knowledge Graph Tutorial

Knowledge Graphs

AIDAN HOGAN, IMFD, DCC, Universidad de Chile, Chile
EVA BLOMQVIST, Linköping University, Sweden
MICHAEL COCHEZ, Vrije Universiteit and Discovery Lab, Elsevier, The Netherlands
CLAUDIA D'AMATO, University of Bari, Italy
GERARD DE MELO, HPI, Germany and Rutgers University, USA
CLAUDIO GUTIERREZ, IMFD, DCC, Universidad de Chile, Chile
JOSÉ EMILIO LABRA GAYO, Universidad de Oviedo, Spain
SABRINA KIRRANE, SEBASTIAN NEUMAIER, and AXEL POLLERES, WU Vienna, Austria
ROBERTO NAVIGLI, Sapienza University of Rome, Italy
AXEL-CYRILLE NGONGA NGOMO, DICE, Universität Paderborn, Germany
SABBIR M. RASHID, Tetherless World Constellation, Rensselaer Polytechnic Institute, USA
ANISA RULA, University of Milano–Bicocca, Italy and University of Bonn, Germany
LUKAS SCHMELZEISEN, Universität Stuttgart, Germany
JUAN SEQUEDA, data.world, USA
STEFFEN STAAB, Universität Stuttgart, Germany and University of Southampton, UK
ANTOINE ZIMMERMANN, École des mines de Saint-Étienne, France

In this paper we provide a comprehensive introduction to knowledge graphs, which have recently garnered significant attention from both industry and academia in scenarios that require exploiting diverse, dynamic, large-scale collections of data. After some opening remarks, we motivate and contrast various graph-based data models and query languages that are used for knowledge graphs. We discuss the roles of schema, identity, and context in knowledge graphs. We explain how knowledge can be represented and extracted using a combination of deductive and inductive techniques. We summarise methods for the creation, enrichment, quality assessment, refinement, and publication of knowledge graphs. We provide an overview of prominent open knowledge graphs and enterprise knowledge graphs, their applications, and how they use the aforementioned techniques. We conclude with high-level future research directions for knowledge graphs.

CCS Concepts: • **Information systems** → **Graph-based database models**; **Information integration**;

Additional Key Words and Phrases: knowledge graph

1 INTRODUCTION

Though the phrase "knowledge graph" has been used in the literature since at least 1972 [465], the modern incarnation of the phrase stems from the 2012 announcement of the Google Knowledge Graph [484], followed by further announcements of the development of knowledge graphs by Airbnb [87], Amazon [298], eBay [417], Facebook [387], IBM [128], LinkedIn [224], Microsoft [482], Uber [214], and more besides. The growing industrial uptake of the concept proved difficult for academia to ignore: more and more scientific literature is being published on knowledge graphs, which includes books (e.g. [425]), as well as papers outlining definitions (e.g., [141]), novel techniques (e.g., [318, 424, 553]), and surveys of specific aspects of knowledge graphs (e.g., [400, 549]).

Underlying all such developments is the core idea of using graphs to represent data, often enhanced with some way to explicitly represent knowledge [387]. The result is most often used in application scenarios that involve integrating, managing and extracting value from diverse sources of data at large scale [387]. Employing a graph-based abstraction of knowledge has numerous benefits in such settings when compared with, for example, a relational model or NoSQL alternatives. Graphs provide a concise and intuitive abstraction for a variety of domains, where edges capture the (potentially cyclical) relations between the entities inherent in social data, biological interactions,

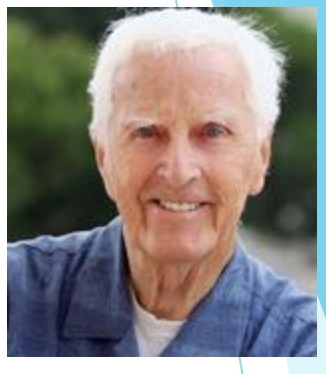
https://arxiv.org/pdf/2003.02320

Recall - Degree Centrality

- Total number of connections a vertex has
  - Degree of vertex
    - Total number of edges connected to a vertex
- Directed network
  - In-degree
  - Out-degree
- Can be considered a popularity measure
  - Is it a good popularity measure?

Recall - Betweenness Centrality

Another way of looking at centrality is by considering how important nodes are in connecting other nodes



Linton Freeman

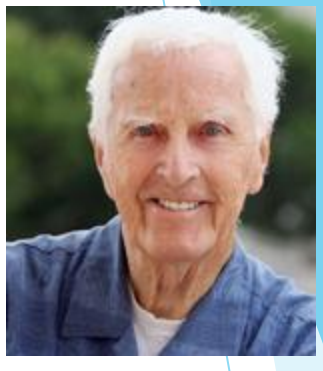
$$C_b(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

σ_{st} The number of shortest paths from vertex s to t – *a.k.a.* information pathways

$\sigma_{st}(v_i)$ The number of **shortest paths** from s to t that pass through v_i

Recall - Closeness Centrality

- The intuition is that influential/central nodes can quickly reach other nodes



Linton Freeman

- These nodes should have a smaller average shortest path length to others

Closeness centrality: $C_c(v_i) = \frac{1}{\bar{l}_{v_i}}$

$$\bar{l}_{v_i} = \frac{1}{n-1} \sum_{v_j \neq v_i} l_{i,j}$$

An Interesting Comparison!

Comparing three centrality values

- Generally, the 3 centrality types will be positively correlated
- When they are not (or low correlation), it usually reveals interesting information

	Low Degree	Low Closeness	Low Betweenness
High Degree		*Node is embedded in a community that is far from the rest of the network*	*Ego's connections are redundant - communication bypasses the node*
High Closeness	*Key node connected to important/active alters*		*Probably multiple paths in the network, ego is near many people, but so are many others*
High Betweenness	*Ego's few ties are crucial for network flow*	*Very rare! Ego monopolizes the ties from a small number of people to many others.*	

This slide is modified from a slide developed by James Moody

Clustering Coefficient

- Captures how connected your friends are
- For each vertex, it is a measure of the density of the 1.5-degree egocentric network
- Value between 0 and 1
- The higher the value, the more an individual's friends know each other

Clustering and Community Detection

- Want to determine existence and boundaries of groups within networks
- Different than components
  - Want to determine groups within components

Eigenvector Centrality

- Can be considered a measure of influence of an individual in a network
- Involves the connections of a vertex as well as the degree of the connected vertices
  - i.e. Friends of friends
- Higher value indicates more influence

Eigenvector Centrality

- Eigenvectors play a significant role in network analysis/machine learning
- Why?
  - Graphs can be represented as adjacency matrices
    - Adjacency matrices are always square
  - Eigenvectors provide information about matrices

Eigenvectors and Eigenvalues

Eigenvalues and Eigenvectors

Here is the most important definition in this text.

⚠ **Definition.** Let A be an $n \times n$ matrix.

1. An ***eigenvector*** of A is a *nonzero* vector v in $\mathbf{R}^n$ such that $Av = \lambda v$, for some scalar λ.
2. An ***eigenvalue*** of A is a scalar λ such that the equation $Av = \lambda v$ has a *nontrivial* solution.

If $Av = \lambda v$ for $v \neq 0$, we say that λ is the ***eigenvalue for*** v, and that v is an ***eigenvector for*** λ.

The German prefix "eigen" roughly translates to "self" or "own". An eigenvector of A is a vector that is taken to a multiple of itself by the matrix transformation $T(x) = Ax$, which perhaps explains the terminology. On the other hand, "eigen" is often translated as "characteristic"; we may think of an eigenvector as describing an intrinsic, or characteristic, property of A.

Note. Eigenvalues and eigenvectors are only for square matrices.

Eigenvectors are *by definition nonzero*. Eigenvalues may be equal to zero.

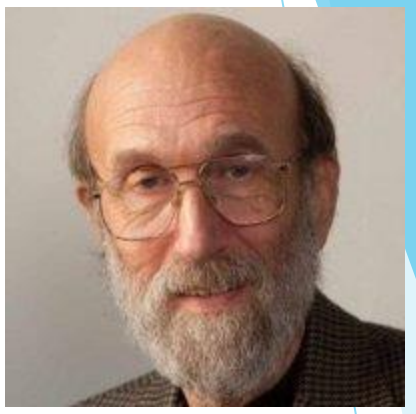
https://textbooks.math.gatech.edu/ila/eigenvectors.html

Eigenvalue Centrality Measures

- https://ocw.mit.edu/courses/14-15-networks-spring-2022/resources/mit14_15s22_lec3/

Eigenvector Centrality

- Having more friends does not by itself guarantee that someone is more important
  - Having more **important friends** provides a stronger signal



Phillip Bonacich

- Eigenvector centrality generalizes degree centrality by incorporating the importance of the neighbors (undirected)

- For directed graphs, we can use incoming or outgoing edges

Katz Centrality

- A major problem with eigenvector centrality arises when it deals with directed graphs

Elihu Katz

- Centrality only passes over *outgoing* edges and in special cases such as when a node is in a directed acyclic graph centrality becomes zero
  - The node can have many edge connected to it
- To resolve this problem we add bias term β to the centrality values for all nodes

Eigenvector Centrality

$$C_{\mathrm{Katz}}(v_i) = \alpha \sum_{j=1}^{\mathrm{n}} A_{j,i} C_{\mathrm{Katz}}(v_j) + \beta$$

PageRank

- Problem with Katz Centrality:
  - In directed graphs, once a node becomes an authority (high centrality), it passes **all** its centrality along **all** of its out-links
- This is less desirable since not everyone known by a well-known person is well-known
- **Solution?**
  - We can divide the value of passed centrality by the number of outgoing links, i.e., out-degree of that node
  - Each connected neighbor gets a fraction of the source node's centrality

PageRank

- Algorithm used by Google (at least historically) to rank web pages within search engine results
- Intuition is that the number and quality of links can provide an estimate of how important the website it

Search Engines

- As a website designer, how do you get your website to be viewed in relevant situations?
  - How can you increase the likelihood that an interested person will find your website?

Search Engine Optimization

Search engine optimization

Search engine optimization is the process of improving the quality and quantity of website traffic to a website or a web page from search engines. SEO targets unpaid traffic rather than direct traffic or paid traffic. Unpaid traffic may originate from different kinds of searches, including image search, video search, academic search, news search, and industry-specific vertical search engines.

Wikipedia

Knowledge Graphs (https://arxiv.org/pdf/2003.02320)

- The term “knowledge graph” has been used since at least 1972
- Popularized in more recent times with Google Knowledge Graph in 2012
- Core idea is using graphs to represent data
  - Often enhanced with a way to explicitly represent knowledge
- Knowledge graphs often used in application scenarios that involve
  - Integrating
  - Managing
  - Extracting value

  From diverse sources of large scale data

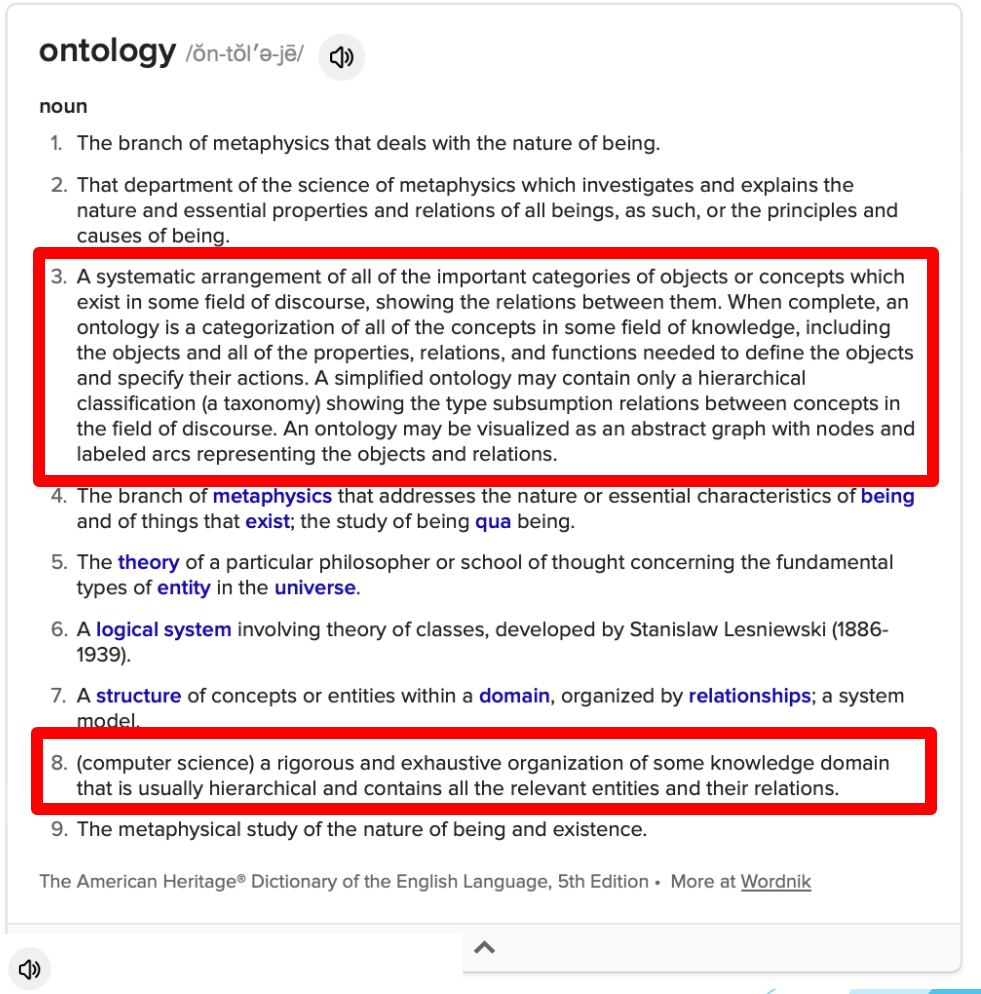
Benefits of Knowledge Graphs

- Concise and intuitive abstraction
- Edge capture relations between entities (nodes)
- Allows for designers to postpone definition of schema
  - Data can evolve in more flexible manner
  - Good for capturing incomplete knowledge
- Graph query languages support traditional relational db operators as well as navigational operators
- **Opens up a range of techniques that can be used for integrating and extracting value from diverse sources of data**

Knowledge graphs as foundation

- Standard knowledge representation formalisms
  - Ontologies
  - Rules

  Can be use to define and reason about the semantics of the terms used to label and describe the nodes and edges in the graph
- Scalable frameworks for graph analytics
  - Can calculate network metrics (centrality, clustering, etc.)
- Machine learning



ontology /ŏn-tŏl′ə-jē/

noun

1. The branch of metaphysics that deals with the nature of being.
2. That department of the science of metaphysics which investigates and explains the nature and essential properties and relations of all beings, as such, or the principles and causes of being.
3. A systematic arrangement of all of the important categories of objects or concepts which exist in some field of discourse, showing the relations between them. When complete, an ontology is a categorization of all of the concepts in some field of knowledge, including the objects and all of the properties, relations, and functions needed to define the objects and specify their actions. A simplified ontology may contain only a hierarchical classification (a taxonomy) showing the type subsumption relations between concepts in the field of discourse. An ontology may be visualized as an abstract graph with nodes and labeled arcs representing the objects and relations.
4. The branch of **metaphysics** that addresses the nature or essential characteristics of **being** and of things that **exist**; the study of being **qua** being.
5. The **theory** of a particular philosopher or school of thought concerning the fundamental types of **entity** in the **universe**.
6. A **logical system** involving theory of classes, developed by Stanislaw Lesniewski (1886-1939).
7. A **structure** of concepts or entities within a **domain**, organized by **relationships**; a system model.
8. (computer science) a rigorous and exhaustive organization of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations.
9. The metaphysical study of the nature of being and existence.

The American Heritage® Dictionary of the English Language, 5th Edition • More at Wordnik

semantics /sĭ-măn′tĭks/

noun

1. The study or science of meaning in language.
2. The competence of a speaker with regard to the interpretation of the meaning of linguistic structures.
3. The study of relationships between signs and symbols and what they represent.

The American Heritage® Dictionary of the English Language, 5th Edition • More at Wordnik

Knowledge Graph Definition (per referenced tutorial)

- **“A graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent relations between these entities.”**
  - The data graph conforms to a graph-based data model, which may be a directed edge-labelled graph, a property graph, etc.
  - By knowledge, we refer to something that is known.
    - Such knowledge may be accumulated from external sources or extracted from the knowledge graph itself
    - Knowledge may be composed of simple statements
      - For example, “Santiago is the capital of Chile”
    - Or quantified statements
      - For example, “all capitals are cities”
    - Deductive and inductive methods may be used to accumulate knowledge
      - For example, by deduction the above examples would add “Santiago is a city”

Examples of graph modeling of knowledge

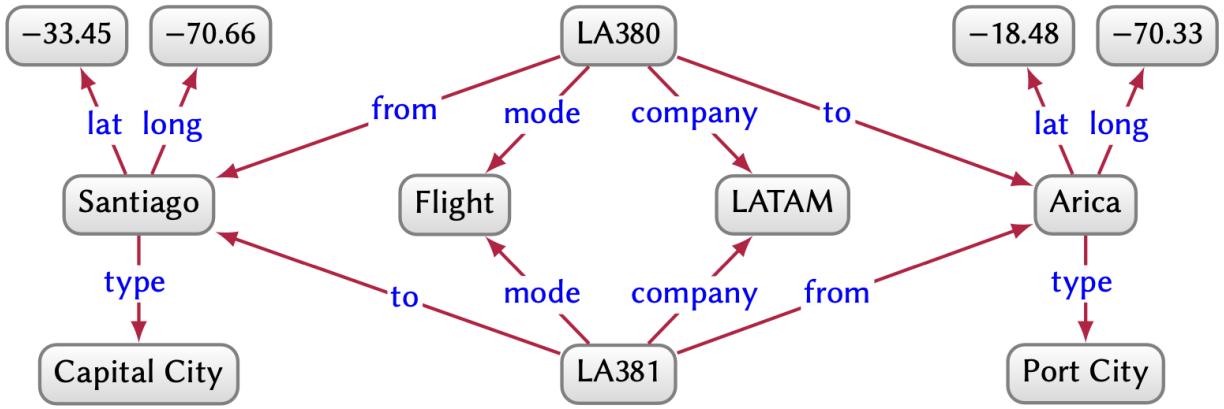


Fig. 3. Directed edge-labelled graph with companies offering flights between Santiago and Arica

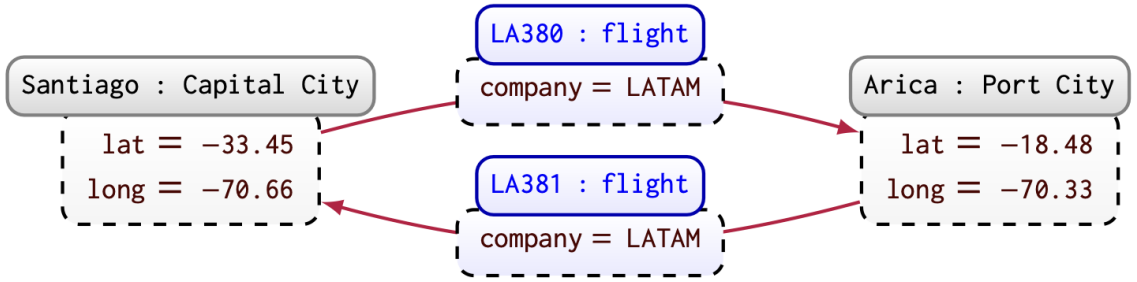
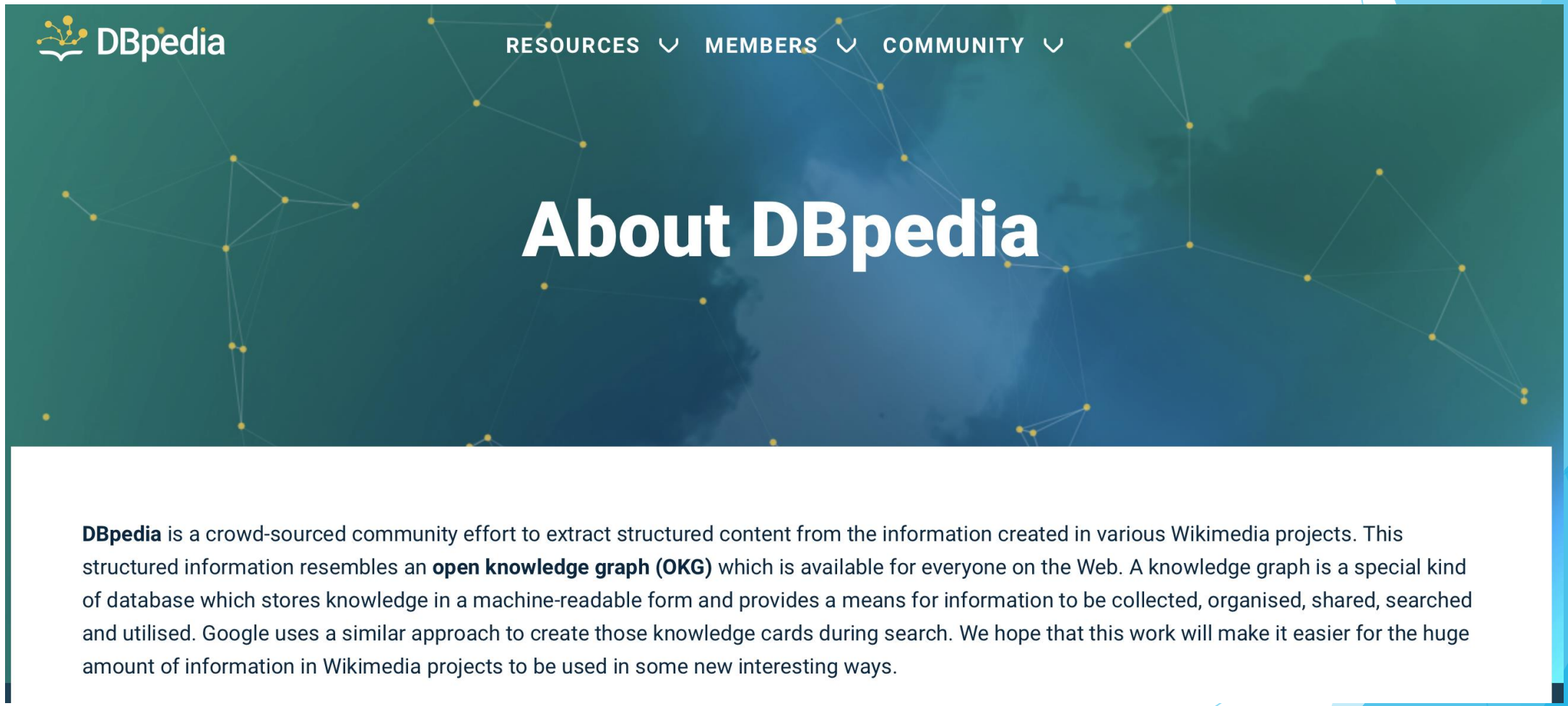


Fig. 4. Property graph with companies offering flights between Santiago and Arica

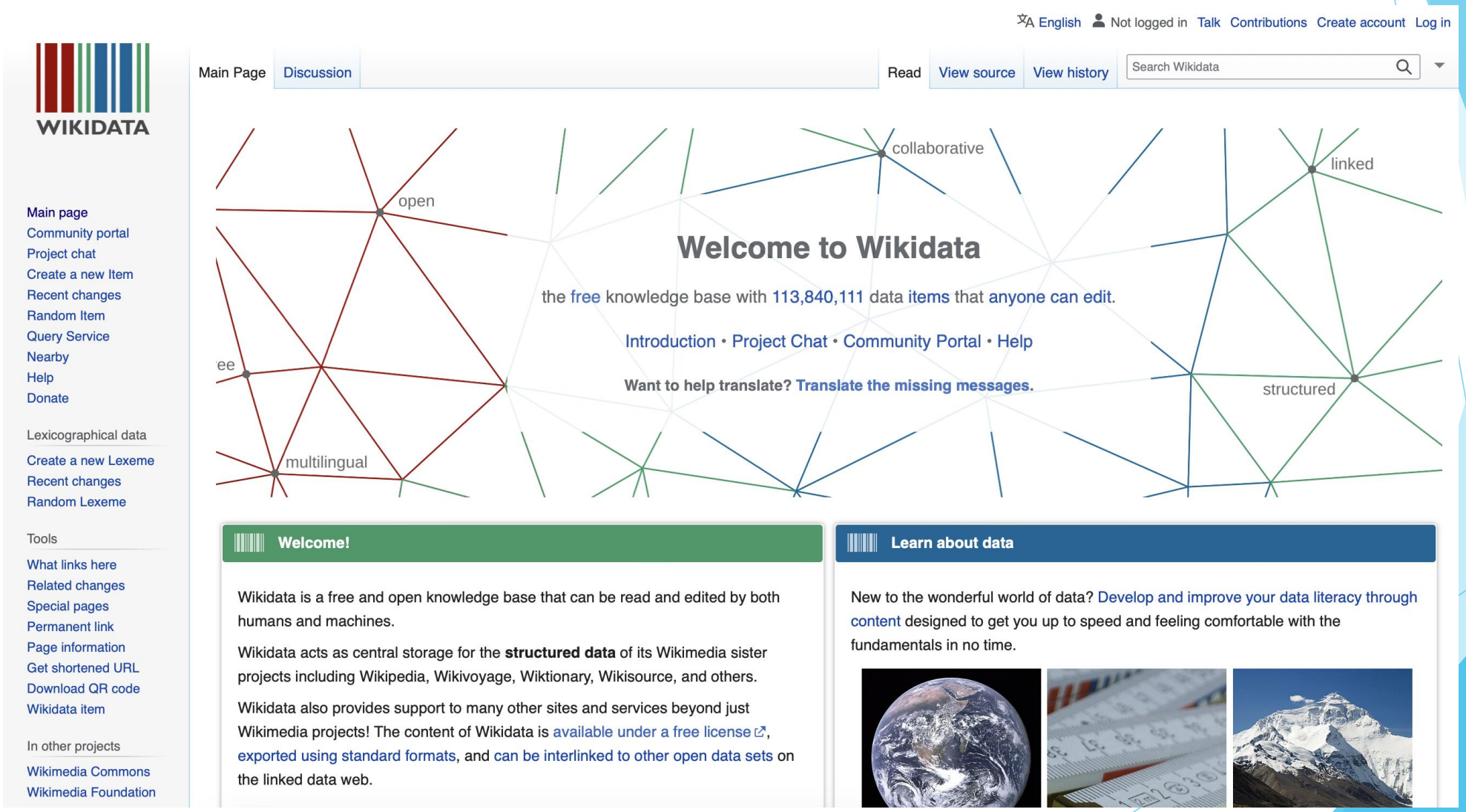
Knowledge graphs in practice

- Shared knowledge within an organization or community
  - Open knowledge graphs
    - Published online
    - Content accessible for public good
    - Cover many domains
    - Many
      - Extracted from Wikipedia
      - Built by communities of volunteers
    - Examples – DBpedia, Wikidata, YAGO

DBpedia.org


DBpedia
RESOURCES
MEMBERS
COMMUNITY
About DBpedia
DBpedia is a crowd-sourced community effort to extract structured content from the information created in various Wikimedia projects. This structured information resembles an open knowledge graph (OKG) which is available for everyone on the Web. A knowledge graph is a special kind of database which stores knowledge in a machine-readable form and provides a means for information to be collected, organised, shared, searched and utilised. Google uses a similar approach to create those knowledge cards during search. We hope that this work will make it easier for the huge amount of information in Wikimedia projects to be used in some new interesting ways.

wikidata.org

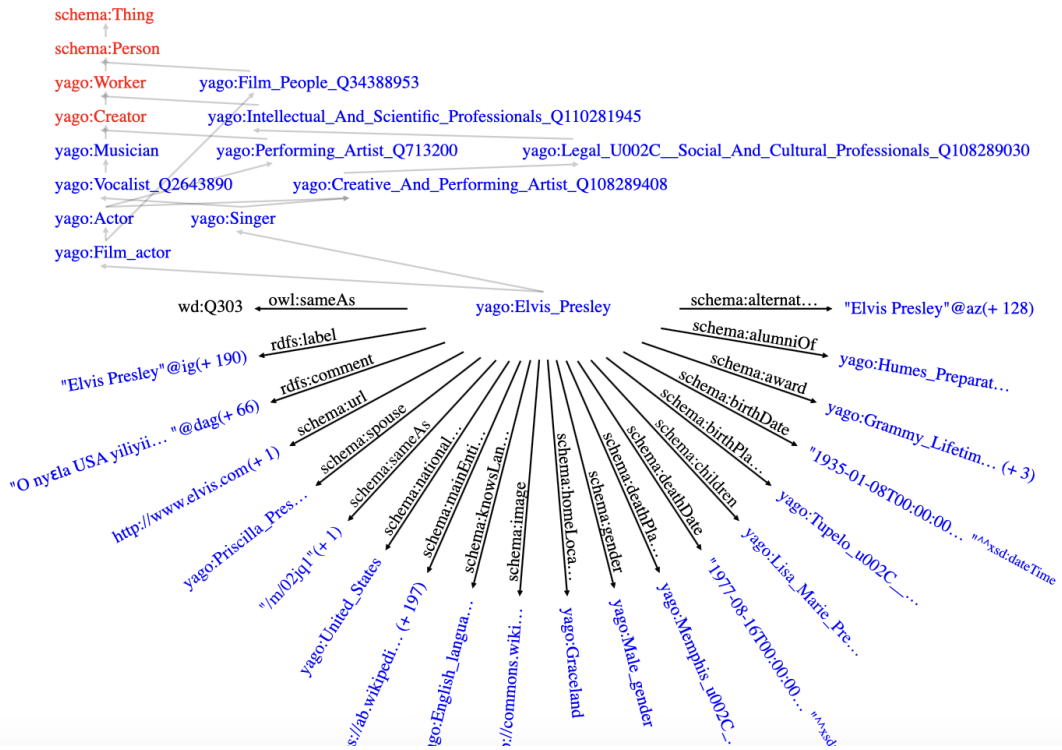


yago-knowledge.org

Home | Getting Started | Downloads | Browse/Query | Publications | Contributors

YAGO: A High-Quality Knowledge Base

YAGO is a large knowledge base with general knowledge about people, cities, countries, movies, and organizations.



Knowledge graphs in practice

- Shared knowledge within an organization or community
  - Enterprise knowledge graphs
    - Typically internal to a company
    - Applied for commercial use-cases, for example
      - Web search
        - Google, Bing
      - Commerce
        - Airbnb, Amazon, eBay, Uber
      - Social Networks
        - Facebook, LinkedIn
      - Finance
        - Accenture, Banca d”Italia, Bloomberg, Capital One, Wells Fargo
    - Examples of applications
      - Search, recommendations, personal agents, advertising, business analytics, risk assessment, automation

See the tutorial for more detail and references

Implementation

- There are many ways to implement knowledge graphs
- Neo4j uses property graphs