

Large Scale Graph Mining

GENERAL INTRODUCTION

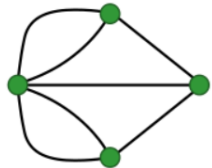
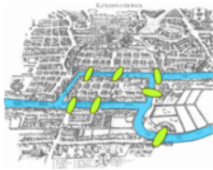
Nacéra Seghouani

Computer Science Department, CentraleSupélec
Laboratoire Interdisciplinaire des Sciences du Numérique, LISN
nacera.seghouani@centralesupelec.fr

2024-2025

The Bridges of Königsberg

- ☞ Königsberg (Kaliningrad), the capital of Eastern Prussia, supported the trade by its fleet of ships by building seven bridges across the river Pregel that surrounded the town.
- ☞ The question is can one walk across all seven bridges and never cross the same one twice?
- ☞ The problem remained unsolved until 1735, when Leonard Euler, a Swiss born mathematician: Euler's proof was the first time someone solved a mathematical problem using a graph.

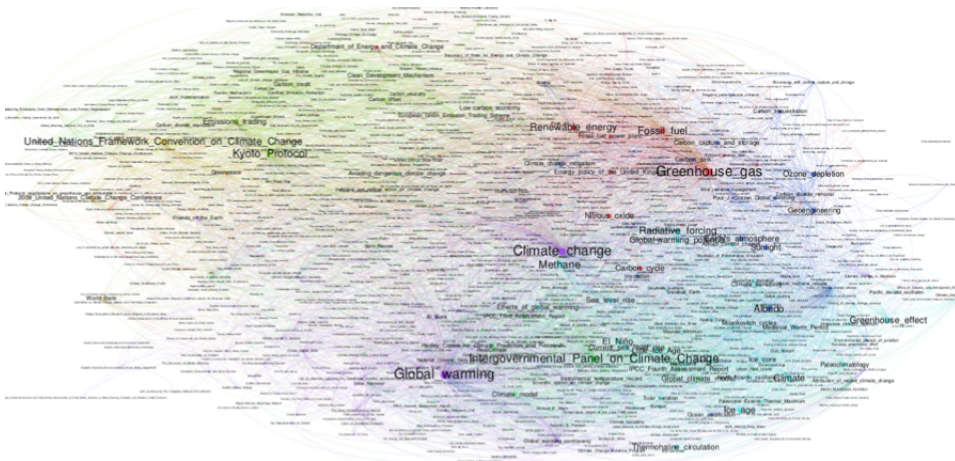


7 Bridges problem ↗

Graph-structured data is everywhere

- ☞ Graph-structured data at the heart of complex systems, play a major role in our daily life, in science and economy
 - ✓ Cooperation between billions of individuals, or communication infrastructures - billions of cell phones with computers and satellites -, interactions between thousands of genes and metabolites within our cells, billions of neurons in our brain, internet of things ...
- ☞ Understanding mathematical foundations, description, prediction, and eventually control this kind of systems is one of the major scientific challenges of the 21st century.
 - ✓ The Bridges of Königsberg (18th century with Euler), Paul Erdős and Alfréd Rényi (1959), Mark Granovetter (1973)
 - ✓ Emergence of Network Science
 - ✓ Also the most revolutionary technologies, empowering everything from Google to Facebook, CISCO, Twitter, LinkedIn, ...
 - ✓ And neural networks ...

- 4/15



Social Networks

- ☞ Vertices: Users, Posts (Videos, Music, Comments ...)
- ☞ Edges: Social Relationships (directed: Twitter, undirected: Facebook), Likes, Texts ...
- ☞ Questions: the influencers, which kind communities and profiles, ...

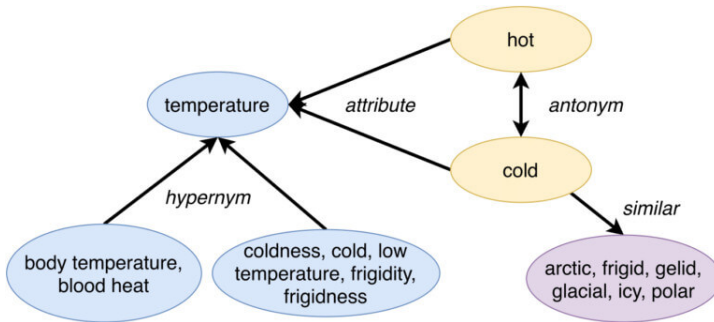


Encyclopedia, Semantic & Knowledge Networks

Wordnet: network of English words (extended to other languages) linked by lexical and semantic relations. It can thus be seen as a combination and extension of a dictionary and thesaurus.

☞ nouns, verbs, adjectives, adverbs, ...

☞ synonymy or the opposite antonymy, meronymy (denotes a constituent part of) or the opposite holonymy, hyponymy (denotes a kind of) or hypernymy ...



Encyclopedia, Semantic & Knowledge Networks

Typical NLP problems

- Information extraction at sentence or document level (across sentences)
- Word sense disambiguation (neighbouring words)
- Entity Resolution, Entity Linking, Relation Extraction

Part-of-speech tagging (POS tagging)

Natural language processing is an area of computer science and artificial intelligence concerned with the interactions between computers and human natural languages

Diagram illustrating Part-of-speech tagging (POS tagging) for the sentence: "Natural language processing is an area of computer science and artificial intelligence concerned with the interactions between computers and human natural languages". The words are tagged with their respective parts of speech: JJ (Adjective), NN (Noun), VBZ (Verb), DT (Determiner), IN (Preposition), CC (Conjunction), JJ (Adjective), NN (Noun), VBN (Verb), IN (Preposition), DT (Determiner), NNS (Noun), IN (Preposition), NNS (Noun), CC (Conjunction), JJ (Adjective), JJ (Adjective), NNS (Noun).

Named entity recognition

Thomas Cruise, born at Syracuse, New York, U.S., July 3, 1962, is an American actor and producer. He has been nominated for three Academy Awards and has won three Golden Globe Awards. He started his career at age 19 in the film Endless Love.

Diagram illustrating Named entity recognition (NER) for the sentence: "Thomas Cruise, born at Syracuse, New York, U.S., July 3, 1962, is an American actor and producer. He has been nominated for three Academy Awards and has won three Golden Globe Awards. He started his career at age 19 in the film Endless Love." The entities are tagged: Person (Thomas Cruise), Location (Syracuse, New York, U.S.), Date (July 3, 1962), Misc (American actor and producer), Number (three), and Num (19).

Dependency Parsing

Natural language processing is an area of computer science and artificial intelligence concerned with the interactions between computers and human natural languages

Diagram illustrating Dependency Parsing for the sentence: "Natural language processing is an area of computer science and artificial intelligence concerned with the interactions between computers and human natural languages". The diagram shows the syntactic structure with dependencies between words and their parts of speech. Key dependencies include: "Natural language" (JJ) to "processing" (NN) via amod; "processing" (NN) to "is" (VBZ) via nsubj; "is" (VBZ) to "an" (DT) via cop; "an" (DT) to "area" (NN) via det; "area" (NN) to "of" (IN) via case; "of" (IN) to "computer" (NN) via nmod; "computer" (NN) to "science" (NN) via compound; "science" (NN) to "and" (CC) via conj; "and" (CC) to "artificial" (JJ) via compound; "artificial" (JJ) to "intelligence" (NN) via compound; "intelligence" (NN) to "concerned" (VBN) via amod; "concerned" (VBN) to "with" (IN) via case; "with" (IN) to "the" (DT) via det; "the" (DT) to "interactions" (NNS) via nmod; "interactions" (NNS) to "between" (IN) via case; "between" (IN) to "computers" (NNS) via nmod; "computers" (NNS) to "and" (CC) via conj; "and" (CC) to "human" (JJ) via compound; "human" (JJ) to "natural" (JJ) via compound; "natural" (JJ) to "languages" (NNS) via compound.

Encyclopedia, Semantic & Knowledge Networks

- ☞ Vertices: Articles
- ☞ Edges: cross-language links, internal links, category links, redirect links ...

WIKIPEDIA

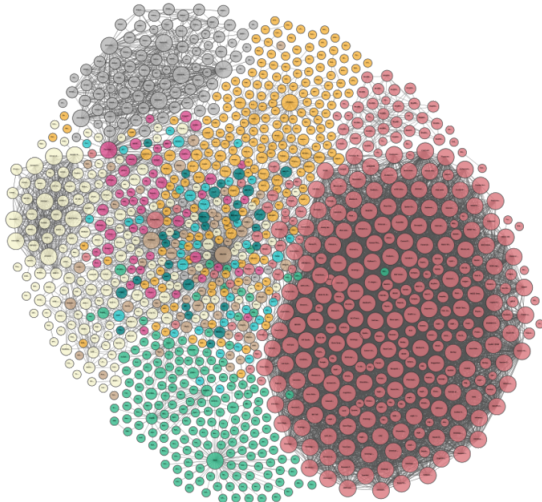
The Free Encyclopedia



Encyclopedia, Semantic & Knowledge Networks

DBpedia: Open data project built from wikipedia and other thesaurus based on ontologies (Yago, wordnet, DBLP, Freebase, Wikidata...). Used by Google, IBM Watson's, ... Popular: Google and Satori Knowledge Graphs [↗](#)

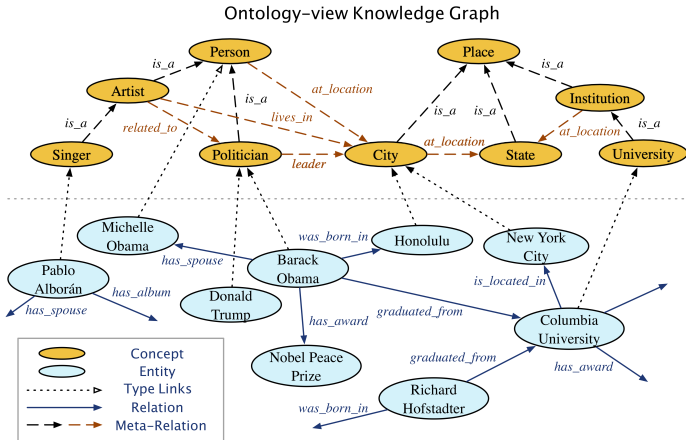
- ✎ According to Gruber, *ontology is an explicit specification of a conceptualization*, which makes it possible to specify in a formal language the concepts of a domain and their relationships.
IBM Watson's Jeopardy!



Encyclopedia, Semantic & Knowledge Networks

☞ Vertices: Subject s , Object o - Edges: Semantic relationships r

☞ KB is a set of triples $\langle s, r, o \rangle$: T-box (is-a, subclassOf, equivalentClass, properties, domain, range, ...) and A-box (class instances and their relationships)

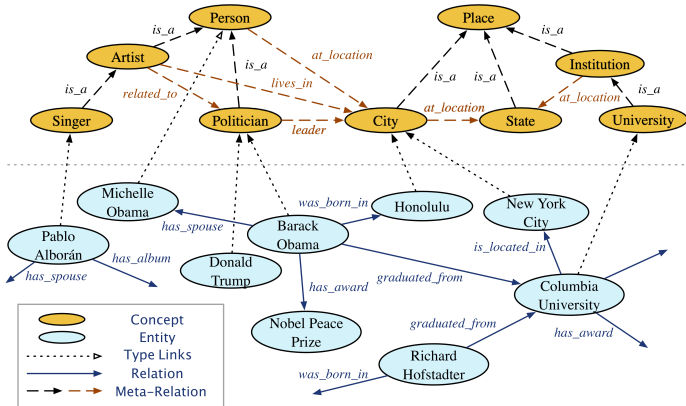


Encyclopedia, Semantic & Knowledge Networks

A knowledge graph:

- ☞ Inference engine based on description logics Challenges: Incomplete, Inconsistent, Ambiguous
- ☞ W3C languages: RDF, OWL, SPARQL, ...

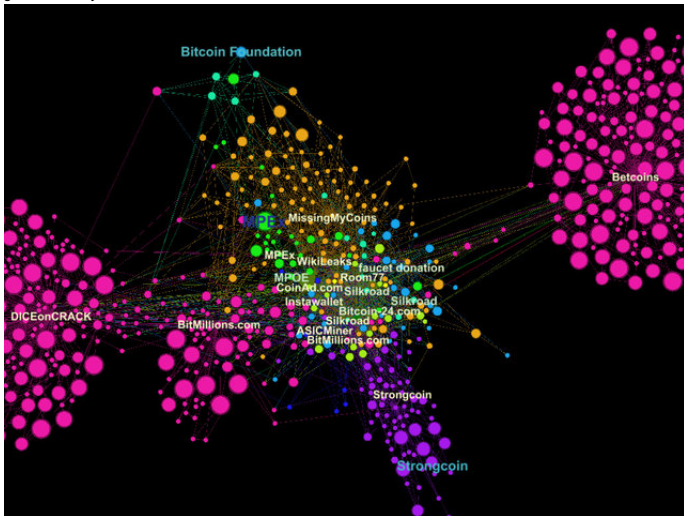
Ontology-view Knowledge Graph



Instance-view Knowledge Graph

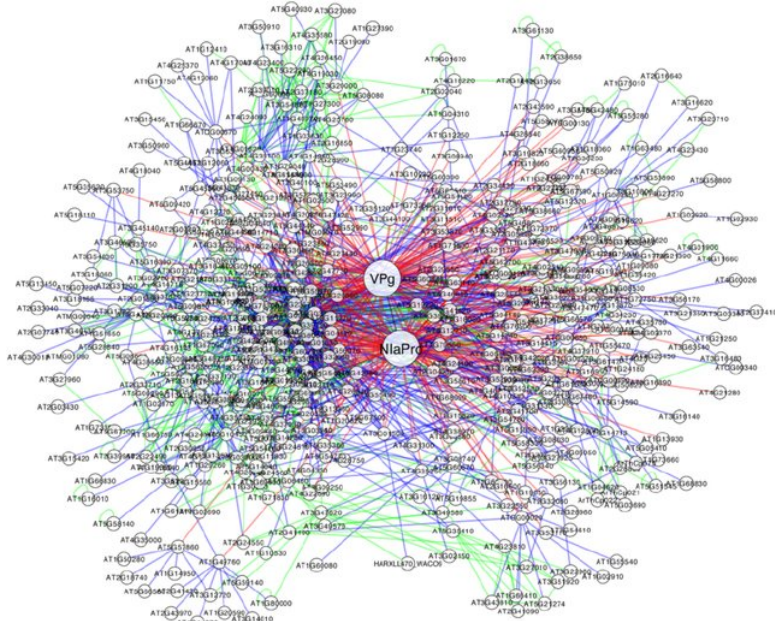
Transaction Networks (e.g., Bitcoin)

- Vertices: Users
- Edges: Exchange of Currency



Biological Networks

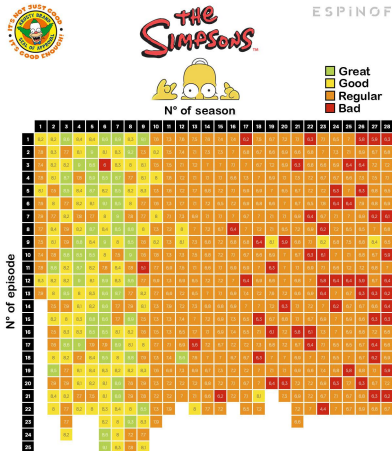
Protein-protein interaction, Metabolic networks, Gene regulatory networks



Recommender Systems

IMDb: Database related to films, television programs, videos, games, streaming content, ...

- Vertices: Users and Items (content-based) or Users and Users (collaboration-based)
- Edges: ratings or similarities



Outline

☞ Main Topics:

PART 1

- ✓ Preliminaries on graph theory & linear algebra
- ✓ Random Walk & PageRank algorithms
- ✓ Graph centrality measures
- ✓ Community detection, Influence maximisation problems

PART 2

- ✓ Embedding graphs: Spectral approaches, Learning-based graph embedding approaches.

☞ Assessment: different kind of grading during the classes

- ✓ Assessment: Mini-Project (deliver + presentation/individual questions).
- ✓ No resit exam.