

Romain Fontugne - iijlab seminar – 2024/10/29

### AGENDA

Why knowledge graphs? Popular knowledge graphs Example applications IYP: Knowledge graph for the Internet

# DATA, DATA, DATA (AND METADATA)

- Data collection everywhere!
- From different sources: application-driven, monitoring, tracking, survey, experiments
- In different forms:
   Database, cloud, data store / lake / silos / warehouse

### $\rightarrow$ Data is valuable, make good use of yours

 $\frac{\text{The}}{\text{Economist}} \equiv Menu \quad Q$ 



### Leaders | Regulating the internet giants

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules



A NEW commodity spawns a lucrative, fast-growing industry, prompting antitrust regulators to step in to restrain those who control its flow. A century ago, the resource in question was oil. Now similar concerns are being raised by the giants that deal in data, the oil of the digital era. These titans—Alphabet (Google's parent company), Amazon, Apple, Facebook and Microsoft—look unstoppable. They are the five most valuable listed firms in the world. Their profits are surging: they collectively racked up over \$25bn in net profit in the first quarter of 2017. Amazon captures half of all dollars spent online in America. Google and Facebook accounted for almost all the revenue growth in digital advertising in America last year.

### **CROSS ANALYSIS**

How to get more from your data?

- Implement data pipelines ingesting different datasets
- Custom-made:
  - Usually serving a single purpose
  - Built for efficiency





### KNOWLEDGE GRAPH

- Generalization of cross analysis
  - Graph-structured data model
  - Semantics (ontology):
    - Nodes / Entities
    - Edges / Relationships



### KNOWLEDGE GRAPH

- Generalization of cross analysis
  - Graph-structured data model
  - Semantics (ontology):
    - Nodes / Entities
    - Edges / Relationships



### $\rightarrow$ Self-explainable data structure!

# KNOWLEDGE GRAPH: DEFINITIONS

- From Wikipedia: "There is no single commonly accepted definition of a knowledge graph."
- Entities: objects, events, or abstract concepts
- Relationships
- Facts (triples): <Da Vinci, painted, Mona Lisa>



### CREATING A KNOWLEDGE GRAPH

Øntology: Naming, hierarchy, level of detail

"there is no single correct ontology for any domain"

- Knowledge Acquisition
  - Entity/Relation extraction
  - Attribute extraction

Knowledge Fusion

•

 Entity Alignment: John Smith = J. Smith?

Disagreeing datasets?

Task Pipeline Integrated Knowledge Graph Versions Un-, Semi- or Structured Input Sources (+ KG and Configs) Knowledge Extraction Ontology Managemen Configurations (Schemas, Mappings) S\_n S\_1 Entity Data Acquisition &  $\mathbb{A}$ Resolution Provenance MetaData Versions KG\_1 KG n & Fusion Preprocessing Repository KG\_n-1 Configs Metadata Management Assurance Knowledge Completion (Cleaning, Mapping) 8 https://www.mdpi.com/2078-2489/15/8/509

### EXAMPLE KNOWLEDGE GRAPH

	Entities	Facts	Comments
Google (2022)	5 billion	500 billion	Google search
Microsoft (2019)	2 billion	55 billion	Bing, Academic, LinkedIn
Facebook (2019)	50 million	500 million	Rebuilt every day
Wikidata (2023)	100 million	15 billion	+10k relationship types manually curated

And a lot more: Netflix, Amazon, eBay, IBM, NASA, ...



## INFORMATION RETRIEVAL

- Example: Google Search
- "things, not strings"
  - Find the right thing
  - Get the best summary
  - Go deeper and broader

Also allow exploratory search

### X 🌷 💽 🔍 Tools All Images News Shopping Videos Maps Web : More インターネットイニシアティブ-IIJ https://www.iij.ad.jp · Translate this page トップ|インターネットイニシアティブ (IIJ) 株式会社インターネットイニシアティブ(IIJ)のオフィシャルサイト。IIJはクラウドサービス からインターネット接続サービス、セキュリティサービス、アウトソーシング. Q Results from iii.ad.ir 会社概要 本社 - 沿革 - 組織図 - ... 個人のお客様 IIJmio (アイアイジェイミオ)は、格安SIMや最新スマホを取り扱う個人向... Internet Initiative Japan Inc IIJ will leverage its outstanding tech 採用情報 From 「採用情報」を掲載しています IIJについて Google IIJの強み - 沿革 - 組織図 - ... **Questions & answers** Knowledge See all questions (1) HIJmio https://www.iijmio.jp · Trans Graph! Reviews <sup>①</sup> 格安SIM/格安スマホ・イ IIJmio (アイアイジェイミオ) は、 ービスです。話題のeSIMや5G対応を 会員専用ページ・ログイン・端末ラインアップ・料金表 IIJmio https://www.iijmio.jp > member - Translate this page : 会員専用ページ IIJmioの会員専用ページです。ログインいただくことで、ご利用料金、データ残量、サービス利 用状況の確認やご契約内容の変更などご利用いただけます。 People also ask Is the IIJ network good in Japan? What does IIJ do? V

Google

What is IIJ Japan?

What are the plans of Internet Initiative Japan?



100 Google reviews Profiles  $\mathbb{X}$ YouTube Facebook X (Twitter People also search for View 15+ more (株)デジタル 株式会社イ **eBOOK** ンターネッ イノベー Initiative プ東京本社 ションミラ トイニシ... ションイ... Japan Co.. (㈱経営... -1(株) Software Software Corporate Corporate Software company office company About this data

× ×

Feedback

Ask a question

Write a review Add a photo

## SEMANTIC SEARCH

- Looking for semantic patterns, not keywords!
- Include context & user intent



## QUESTION ANSWERING

- Example: IBM Watson, Siri, OK Google
- Graph RAG: Help LLMs by giving them context



## **REASONING: GRAPH TRAVERSAL**

- Based on explicit knowledge
  - Deduce new knowledge
  - Identify wrong knowledge



## **REASONING: EMBEDDING**

- Example: Recommendation systems Netflix, Amazon, Facebook
- The graph is too constrained, project data in a latent space





## EMBEDDING: APPLICATIONS

### Helpful for numerous tasks:

- Similarity
- Clustering
- Classification
- Link prediction / recommendation



### EMBEDDING: ALGORITHMS

How to find good embeddings?

• Learning from facts: <h, r, t>

Model	Scoring Function
TransE	$\ \boldsymbol{h}+\boldsymbol{r}-\boldsymbol{t}\ _i$
TransH	$\ (\boldsymbol{h} - \boldsymbol{w}_r^{\mathrm{T}}\boldsymbol{h}\boldsymbol{w}_r) + \boldsymbol{r} - (\boldsymbol{t} - \boldsymbol{w}_r^{\mathrm{T}}\boldsymbol{t}\boldsymbol{w}_r)\ _i$
TransR	$\ \boldsymbol{W}_r\boldsymbol{h}+\boldsymbol{r}-\boldsymbol{W}_r\boldsymbol{t}\ _i$
RESCAL	$h^{\mathrm{T}} W_r t$
DistMult	$h^{\mathrm{T}}diag(r)t$
ComplEx	$Re(\boldsymbol{h}^{\mathrm{T}}diag(\boldsymbol{r})\boldsymbol{\bar{t}})$

### e.g. <Bob, likes, Star Wars 1>



# DOMAIN SPECIFIC

- e.g. Medical applications
  - DrugBank
  - PharmGKB
  - Gene ontology



### INTERNET YELLOW PAGES

Knowledge graph for Internet:

- 40+ datasets (PeeringDB, CAIDA, RIPE, APNIC, Cloudflare, OONI, BGPKit, BGP.Tools, IHR, ...)
- 12+ million nodes
- 82+ million edges
- Produced every week
- Available online at: https://iyp.iijlab.net





# IYP: INFORMATION RETRIEVAL

- <u>https://ihr.iijlab.net</u>
  - Simple search
  - Predefined templates



# AS2497 - Internet Initiative Japan

		vveek	ly report			
MONITORING	ROUTING	DNS	PEERING	REGISTRATION	RANKINGS	CUSTO
	Top Rank	(		Popular Hostnames	External Lin	ks
<u>ipan</u> KPs in 7 Countries				<u>hotpepper.jp</u> jalan.net	<u>bgp.he.net</u> <u>bgp.tools</u>	
174 IPv4 and 8 IPv6 Originated Prefixes       #2 in IHR country ranking: Total AS (JP)         334 Connected ASes       #2 in IHR country ranking: Total AS (JP)		<u>suumo.jp</u> carsensor.net	<u>peeringdb.com</u> radar.cloudflare.com			
	MONITORING          pan         (Ps in 7 Countries         Pv6 Originated Prefixes         ASes         www.iii.ed.in.(on.()	MONITORING ROUTING Top Rank pan (Ps in 7 Countries Pv6 <u>Originated Prefixes</u> #2 in IHI ASes umunii ad in (cn (	MONITORING ROUTING DNS Top Rank Pan (Ps in 7 Countries Pv6 Originated Prefixes ASes unwriti ed is (as (	MONITORING ROUTING DNS PEERING Top Rank Pan (Ps in 7 Countries Pv6 Originated Prefixes #2 in IHR country ranking: Total AS (JP) ASes www.ii ed in (en (	Weekly report         MONITORING       ROUTING       DNS       PEERING       REGISTRATION         Top Rank       Popular Hostnames         pan       hotpepper.jp       jalan.net         VP6 Originated Prefixes       #2 in IHR country ranking: Total AS (JP)       suumo.jp         ASes       carsensor.net       102/tare orm	Weekly report         MONITORING       ROUTING       DNS       PEERING       REGISTRATION       RANKINGS         Top Rank       Popular Hostnames       External Lin         pan       hotpepper.jp       bgp.he.net         (Ps in 7 Countries       #2 in IHR country ranking: Total AS (JP)       suumo.jp       peeringdb.acarsensor.net         ASes       umwiji ed ip (on (       1002 tran comp       tat sin a radar.cloud

Tags

Carrier



### https://iyp.iijlab.net

• AS member of an IXP?

MATCH (a:AS)-[:MEMBER\_OF]-(ix:IXP)

RETURN COUNT(DISTINCT a)

• Prefixes originated by multiple ASes

MATCH (a:AS)-[:ORIGINATE]-(p:Prefix)-[:ORIGINATE]-(b:AS) WHERE a <> b RETURN COUNT(DISTINCT p)



# IYP: REASONING (GRAPH TRAVERSAL)

### Comments on DNS Robustness, Mark Allman, IMC'18

### Summary

### Approach

This paper investigates the robustness of the DNS ecosystem for popular domain names. Using  $r \ln [2]$ : for the .com, .net, .org TLDs.

#### Datasets

Nine years of data (2009 to 2018):

Alexa top 1M
TLD Zone Files (.com, .net, .org)

Also includes traceroutes (only for 2018).

#### Limitations / Future work

**Only 3 TLDs:** They have only have .com, .net, .org zone files so limit their study to these three TI looking at more TLDs is left for future work (end of 'Dataset A', section 3.1).

**Topological determination**: The topological diversity of servers is checked simply by looking if r or not. The paper says that better historical routing information will be used in future work to re step 3).

Anycast prefixes: One limitation of the original study is to ignore anycast prefixes (they keep th that for them :)

IPv6: Original paper looks only at IPv4? we can do both

### (Section 3.1) Coverage of .com, .net, .org in popularity li

The paper considers domain names from only three TLDs but shows that it represents the major

#### Original results



### https://github.com/InternetHealthReport/iyp-notebooks

DNS robustness

RPKI deployment

(HotNets'15)

(IMC'18)

**IYP Results** 

### # Setup access to IYP

from neo4j import GraphDatabase, RoutingControl
from collections import defaultdict

# Using IYP local instance # URI = "neo4j://localhost:7687" # Using IYP public instance URI = "neo4j+s://iyp-bolt.iijlab.net:7687" AUTH = ('neo4j', 'password') db = GraphDatabase.driver(URI, auth=AUTH)

# Get the percentage of .com, .net, and .org domain names in Tranco top 1M
query = """MATCH (r:Ranking (name:'Tranco top 1M')}-[:RANK]-(d:DomainName)-[:MANAGED\_BY]-(a:AuthoritativeNameServer)
WHERE d.name ENDS WITH '.com' OR d.name ENDS WITH '.net' OR d.name ENDS WITH '.org'
RETURN COUNT(DISTINCT d.name)"""

res, \_, \_ = db.execute\_query(query, database\_="neo4j"); nb\_sld = res(0)[0] print(f'(lu0\*nb\_sld)/2000000:.1f)% of Tranco top1M domain names are under the .com, .net, or .org TLD.')

49.1% of Tranco top1M domain names are under the .com, .net, or .org TLD.

```
# Find the percentage of domain names that have nameservers not in the .com, .net, and .org TLDs
query = ""WATCH (r:Ranking (name: 'Tranco top 1W')-[:RaNK].old:ObmainName].fm:NMAGED_BW (reference_name:'openintel.dr
WHERE (d.name ENDS WITH '.com' OR d.name ENDS WITH '.net' OR d.name ENDS WITH '.org')
WITH d, COLLECT(a) AS ns, COLLECT(m) AS managed
// check if all nameservers are outside the zone and have no glue
WHERE all(a in ns WHERE NOT a.name ENDS WITH '.com' AND NOT d.name ENDS WITH '.net' AND NOT d.name ENDS WITH '.org')
RETURN COUNT(DISTINCT d.name)"**
```

res, \_, \_ = db.execute\_query(query, database\_=\*neo4j"); nb\_excluded = res[0][0] print(f'(100\*nb\_excluded/nb\_sld:.1f)% of Tranco top1M domain names are ignored by the original paper assumptions (only

10.3% of Tranco top1M domain names are ignored by the original paper assumptions (only .com, .net, .org nameservers).

### (Section 4.1) Nameserver Replicas

The paper checks nameserver requirements for each .com, .net, and .org SLD, that is at least two nameservers should be deployed in two different locations (different /24 prefixes).

### **Original Results**



## IYP: QUESTION ANSWERING SYSTEM



### NEXT STEPS

- Recommendation systems
  - Peering recommendations
    - In a specific region?
    - For a specific industry?
  - AS classification
  - Country similarities



# CONCLUSION





## IYP: EXPLORATORY SEARCH



# LLM VS. KG

# Knowledge Graphs (KGs)

### Cons:

- Implicit Knowledge
- Hallucination
- Indecisiveness
- Black-box
- Lacking Domainspecific/New Knowledge

### Pros:

- Structural Knowledge
- Accuracy
- Decisiveness
- Interpretability
- Domain-specific Knowledge
- Evolving Knowledge

### Pros:

- General Knowledge
- Language Processing
- Generalizability

### Cons:

- Incompleteness
- Lacking Language
   Understanding
- Unseen Facts

# Large Language Models (LLMs)





## COMMON CHARACTERISTICS

- Large: Millions or billions of nodes & edges
- Coverage: Usually incomplete
- Correctness: how to resolve disagreeing datasets?
- Freshness: Depend on the kind of information

### SEMANTICS

