



2021-09-23



1. Qualcuno saprebbe riconoscere chi sono le persone in questa slide?
wait 5 seconds
2. Vediamolo:
This 00:10 | All 00:10 | Go to next slide



2021-09-23



1. Putin, Assad, Poroshenko, il re Salman, Ahmadinejad...
Tutti politici di altissimo livello...
 2. che apparvero nel...
- This 00:10 | All 00:20 | Go to next slide**



2021-09-23



1. ...Panama Papers leak.
2. Ma cosa erano i Panama Papers?

This 00:15 | All 00:35 | Go to next slide

MOSSACK FONSECA

[John Doe] Hello. This is John Doe.
Interested in data? [Süddeutsche Zeitung]

We're very interested.

[John Doe] There are a couple of conditions. My life is in danger.
We will only chat over encrypted files.
No meeting, ever.
The choice of stories is obviously up to you.

[Süddeutsche Zeitung] Why are you doing this?

[John Doe]

I want to make these crimes public.

2021-09-23



1. – In aprile 2016,
– grazie ad un informatore ...

This 00:5 | All 00:40 | Go to next slide



2021-09-23



1. – ... l'inchiesta di 307 giornalisti da 76 paesi ha portato alla luce documenti e informazioni
– su miliardi di denaro dirottati da studi legali internazionali e banche verso paradisi fiscali
– per conto di...

[This 00:15 | All 00:55 | Go to next slide](#)



2021-09-23



1. ... leader politici, criminali ...

[This 00:05 | All 01:00 | Go to next slide](#)



2021-09-23

1. ... funzionari d'intelligence, artisti ...

[This 00:05](#) | [All 01:05](#) | [Go to next slide](#)



2021-09-23

1. ... VIP dello sport ...

[This 00:05](#) | [All 01:10](#) | [Go to next slide](#)





2021-09-23

1. ... e dello spettacolo.
wait 5 seconds
2. – Quanto era grande il dataset leakato?
– ... come erano strutturati i dati?
This 00:15 | All 01:25 | Go to next slide



The scale of the leak

Volume of data compared to previous leaks

1,7 GB
Cablegate/Wikileaks (2010)

≈ 2,6 TB
Panama Papers/ICIJ (2016)

■ = 1 GB

260 GB
Offshore Leaks/ICIJ (2013)



4 GB
Luxemburg Leaks/ICIJ (2014)

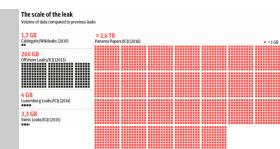


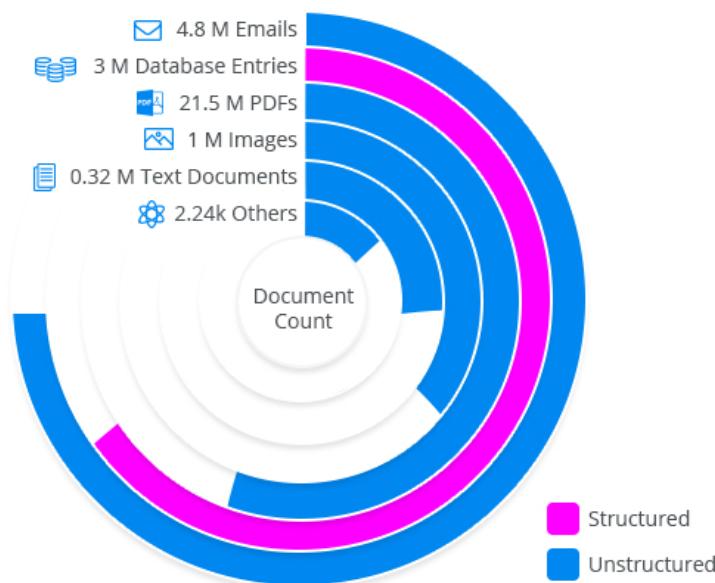
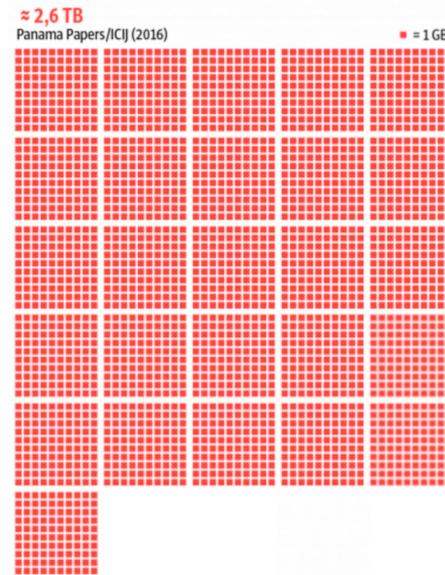
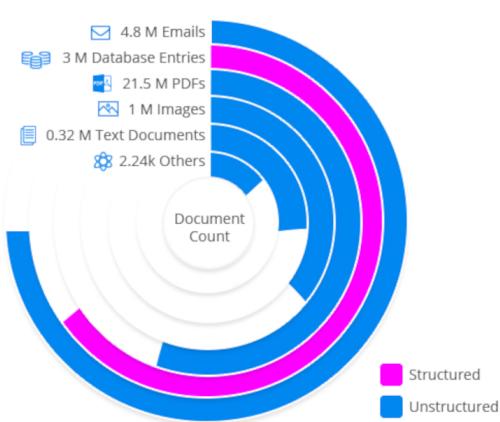
3,3 GB
Swiss Leaks/ICIJ (2015)



2021-09-23

1. Il leak era di dimensione 2.6 TB, 1500 volte più grande del Cablegate del 2010 di Wikileaks.
This 00:10 | All 01:35 | Go to next slide





2021-09-23

A donut chart and a summary table for the dataset. The donut chart shows the distribution of structured vs unstructured data. The summary table provides the following details:

Category	Count
4.8 M Emails	4.8 M
3 M Database Entries	3 M
21.5 M PDFs	21.5 M
1 M Images	1 M
0.32 M Text Documents	0.32 M
2.24k Others	2.24k
Total Document Count	≈ 2.6 TB

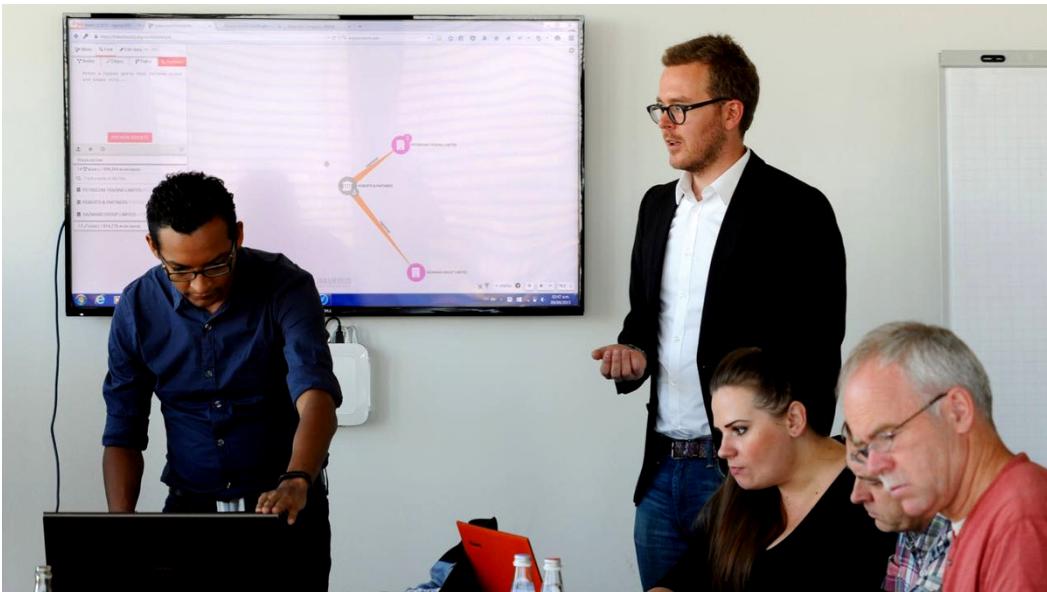
1. La maggior parte dei dati erano non strutturati, sotto forma di emails, immagini e files PDF inviati.
This 00:10 | All 01:45 | Go to next slide

2021-09-23

A donut chart and a summary table for the dataset. The donut chart shows the distribution of structured vs unstructured data. The summary table provides the following details:

Category	Count
4.8 M Emails	4.8 M
3 M Database Entries	3 M
21.5 M PDFs	21.5 M
1 M Images	1 M
0.32 M Text Documents	0.32 M
2.24k Others	2.24k
Total Document Count	≈ 2.6 TB

1. – Una domanda sorge spontanea:
– Come si può indagare su un dataset così immenso
– e in gran parte fatto di dati non strutturati?
This 00:15 | All 02:00 | Go to next slide

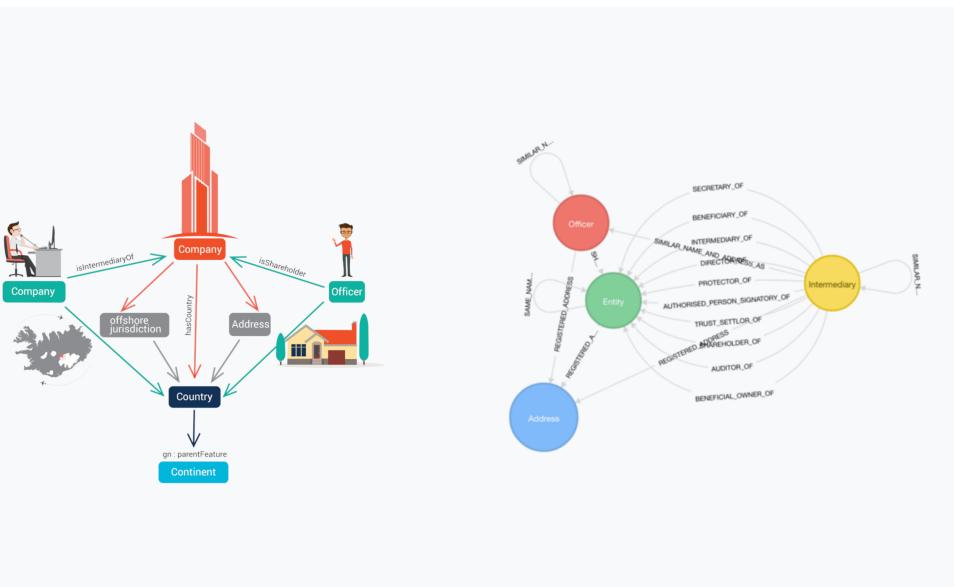


2021-09-23



1. Loro, un team composto da giornalisti investigativi ed esperti di database a grafi ce l'hanno fatta.

This 00:10 | All 02:10 | Go to next slide



2021-09-23



1. Facendo buon uso delle relazioni che derivano da dati interconnessi (ad esempio le mail, le dipendenze tra le entità) ...

This 00:10 | All 02:20 | Go to next slide



```

MATCH (:Address)--(:Officer)--(e:Entity)
MATCH (e)--(:Intermediary)--(:Address)
MATCH (e)--(:Address)
WHERE o.name CONTAINS 'Emma Watson'

```

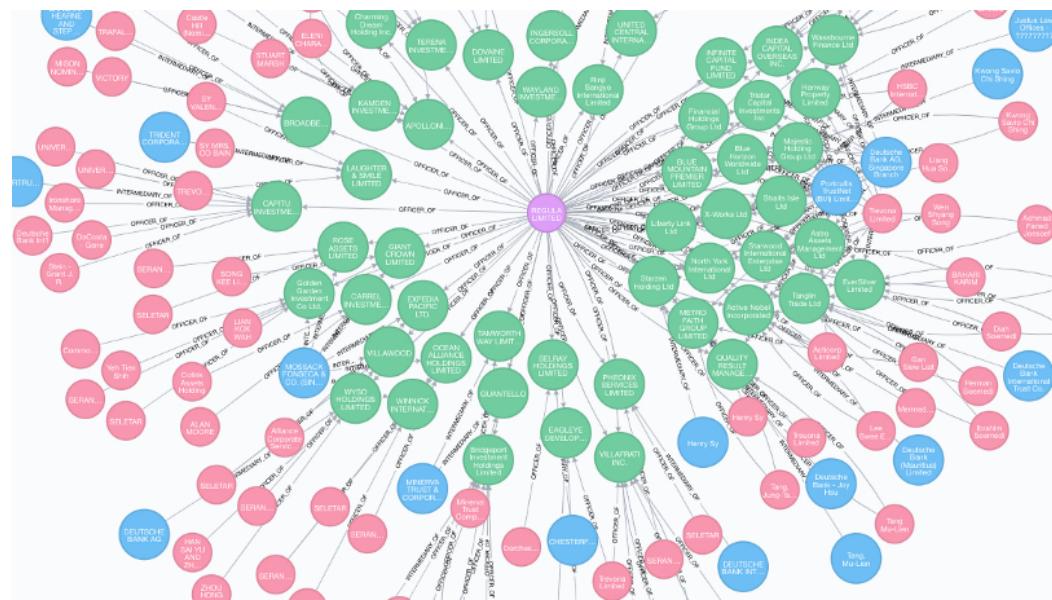
(:Address)--(:Officer)--(:Entity)--(:Intermediary)--(:Address)

2021-09-23

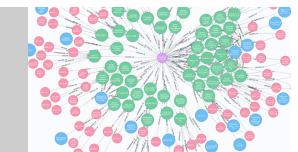


- ... e di nuove tecnologie (come i database a grafi),
 - hanno generato grafi con nodi rappresentanti le entità coinvolte
 - e archi rappresentanti le connessioni tra loro,

This 00:15 | All 02:35 | [Go to next slide](#)



2021-09-23



- il tutto interrogabile
- Questo è uno dei più eclatanti esempi a queste dimensioni
 - dell'utilizzo di database a grafi per svolgere un compito
 - che altri tipi di database semplicemente non potevano fare
 - (almeno i tempi di interrogazione non sarebbero stati umanamente accettabili)
 - perché non progettati per effettuare visite su grafi con dati non strutturati.
- Nel lavoro svolto per la tesi, non si è parlato di Panama Papers.
 - Ma:

This 00:35 | All 03:10 | [Go to next slide](#)

CLUSTERING GRAPHS

Applying a Label Propagation Algorithm to
Detect Communities in Graph Databases

Name Surname



The work done

1. **Literature review** of graph theory, graph databases and clustering algorithms.
2. **dblp.org Dataset download, conversion & import** in ArangoDB Graph DBMS.
3. **Data transformations** to obtain vertices, edges and the complete graph.
4. **Community Detection Algorithm application** on the graph for clustering.
5. **Web Application development** to display the results of the clustering.

2021-09-23

CLUSTERING GRAPHS
Applying a Label Propagation Algorithm to
Detect Communities in Graph Databases
Name Surname

1. – Facendo uso di un database a grafi per il salvataggio
 - e l'interrogazione di dati di un dataset di 5.6 milioni di pubblicazioni scientifiche accademiche da 2.8 milioni di ricercatori,
 - applicando al grafo da essi generati un algoritmo di Community Detection,
 - ovvero, di individuazione delle comunità
 - sono state individuate 180 mila Communities.
 - Inoltre, è stata sviluppata una Full-Stack Web Application per visualizzare queste comunità di collaborazione.
2. Più in dettaglio:

This 00:40 | All 03:50 | [Go to next slide](#)

2021-09-23

The work done

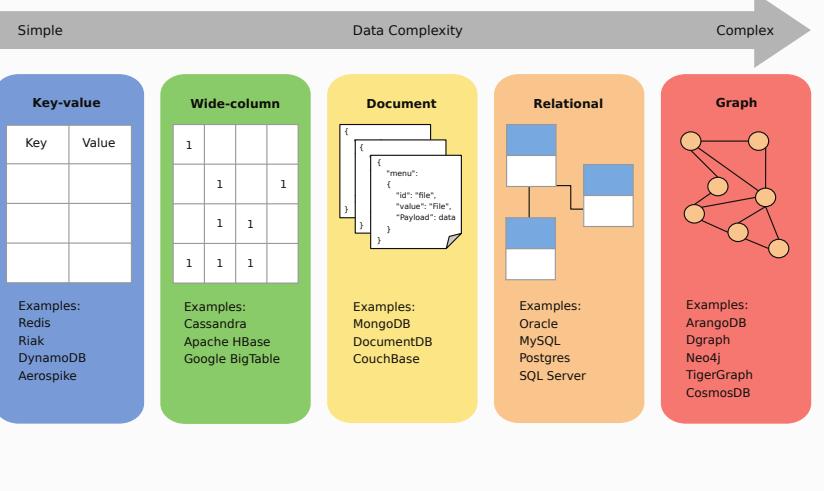
1. Literature review of graph theory, graph databases and clustering algorithms.
2. dblp.org Dataset download, conversion & import in ArangoDB Graph DBMS.
3. Data transformations to obtain vertices, edges and the complete graph.
4. Community Detection Algorithm application on the graph for clustering.
5. Web Application development to display the results of the clustering.

└ The work done

1. – È stata fatta una consultazione in letteratura dei concetti base di teoria dei grafi,
 - delle varie caratteristiche dei database a grafi
 - e degli algoritmi per l'individuazione delle communities in grafi.
2. – Poi è stato scaricato il dataset da dblp.org, è stato convertito
 - e importato su ArangoDB, il graph database management system scelto.
3. – Le collezioni di dati sono state sottoposte a trasformazioni per ottenere vertici, archi e da questi il grafo completo dell'intero dataset.
4. – Su tale grafo poi è stato applicato un Label Propagation Community Detection Algorithm,
 - ovvero un algoritmo a propagazione di etichette per l'individuazione delle comunità di collaborazione scientifica.
5. – In fine, è stata sviluppata una Web Application per visualizzare le communities individuate.
6. Vediamoli uno ad uno...

This 00:50 | All 04:40 | [Go to next slide](#)

Graph databases

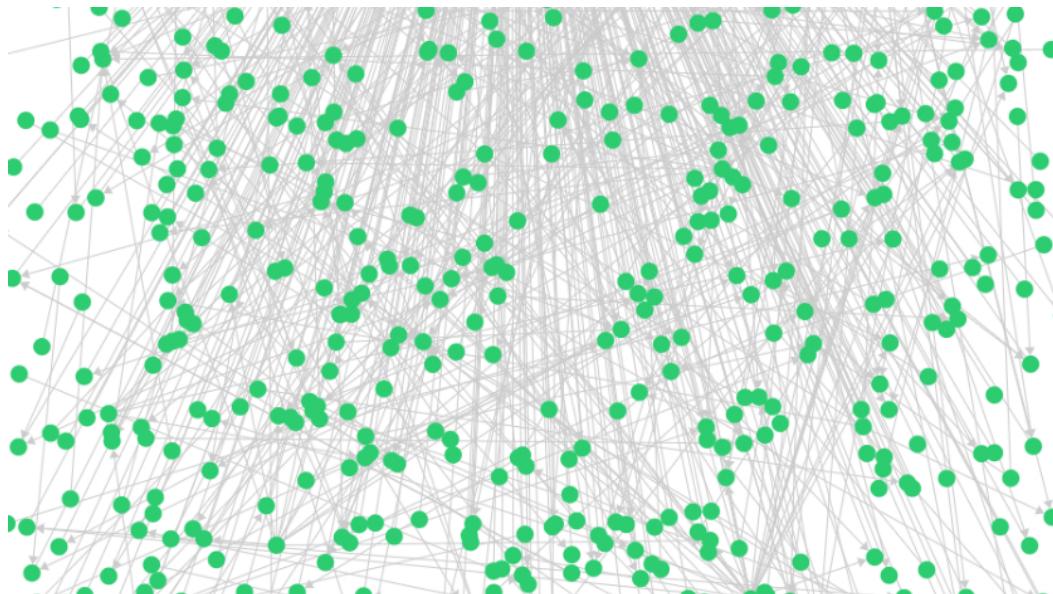


2021-09-23

Graph databases

- ... più i dati sono interconnessi e strutturati in modo complesso,
– più fa senso usare database a grafi
- Inoltre se i nested JOINs da fare durante un'interrogazione sono
più di 3 livelli di profondità,

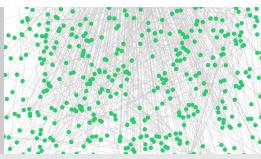
This 00:15 | All 04:55 | [Go to next slide](#)



2021-09-23

- ovvero, in termini di grafi, se bisogna fare più di 3 salti di
attraversamento degli archi,
– allora i tempi di interrogazione usando database tradizionali
diventano proibitivi.
- Dunque, vediamo ora concretamente cosa è stato fatto.

This 00:20 | All 05:15 | [Go to next slide](#)





2021-09-23

└ Le macchine

- Al fine di hostare il database,
 - servire l'API della Web Application
 - e l'interfaccia lato frontend del sito,
 - 3 macchine separate, un router e uno switch sono stati usati.
 - La quarta macchina è stata usata a scopi di sviluppo e accesso remoto.
- Una volta installato i sistemi operativi sulle varie macchine
 - e ArangoDB sulla macchina scelta come host del database...

This 00:30 | All 05:45 | [Go to next slide](#)



The dataset: dblp.org

Name	Last modified	Size	Description
Parent Directory		-	
dblp-2021-09-01.xml.gz.md5	2021-09-01 23:44	57	
dblp-2021-09-01.xml.gz	2021-09-01 23:44	623M	
dblp-2021-08-01.xml.gz.md5	2021-08-02 00:44	57	
dblp-2021-08-01.xml.gz	2021-08-02 00:44	617M	
dblp-2021-07-01.xml.gz.md5	2021-07-02 00:04	57	
dblp-2021-07-01.xml.gz	2021-07-02 00:04	611M	
dblp-2021-06-01.xml.gz.md5	2021-06-02 00:31	57	
dblp-2021-06-01.xml.gz	2021-06-02 00:31	606M	
dblp-2021-05-03.xml.gz.md5	2021-05-03 23:33	57	

The dataset:

- Compressed archive of 623 MB.
- Once extracted: Single XML file of 3.2 GB.
- 8.5 million XML entries on publications, authors, journals, institutions, citations etc.

2021-09-23

└ The dataset: dblp.org

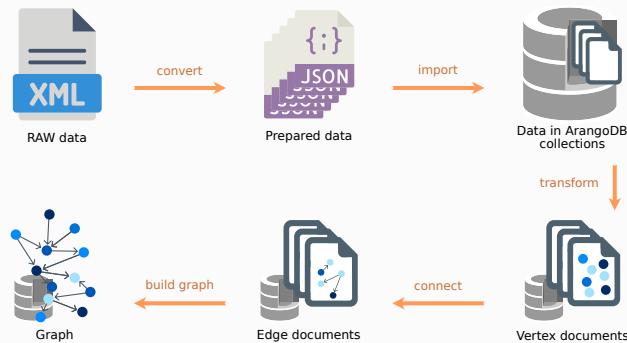
- si procede al download del dataset dal sito di dblp.org .
 - Il dataset è fatto da un unico file compresso di dimensione 623 MB.
 - Una volta estratto, il singolo file XML è di 3.2 GB
 - e contiene circa 8 milioni e mezzo di voci XML su autori, pubblicazioni, affiliazioni, citazioni, journals e così via.
- Sicuramente i dati non possono essere importati nel database così
 - perché ArangoDB richiede che essi siano in formato line JSON.
 - Perciò serve convertire le voci XML in righe di line JSON.
- Un altro problema è la dimensione da 3.2 GB del file XML,
 - non molto facile da manipolare.

This 00:40 | All 6:25 | [Go to next slide](#)



The dataset: dblp.org
Compressed archive of 623 MB
Once extracted: Single XML file of 3.2 GB.
8.5 million XML entries on publications, authors, journals, institutions, citations etc.

Data conversion, import and transformations

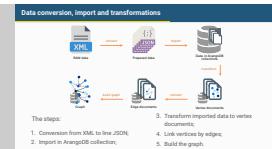


The steps:

1. Conversion from XML to line JSON;
2. Import in ArangoDB collection;
3. Transform imported data to vertex documents;
4. Link vertices by edges;
5. Build the graph.

2021-09-23

└ Data conversion, import and transformations



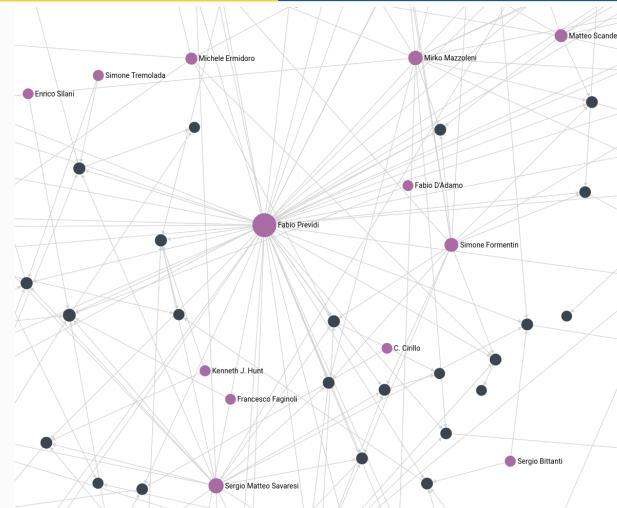
Quindi le operazioni da effettuare per arrivare dai dati grezzi al grafo completo sono:

1. – La suddivisione del file XML in tanti piccoli files
– e la conversione di ciascuno di essi in formato line JSON.
2. L'import dei line JSON in una collection in un database ArangoDB,
3. – La trasformazione dei documenti importati
– in modo che possano essere a tutti gli effetti dei vertici di un grafo.
4. – La creazione degli archi tra i vertici usando gli attributi `_from` e `_to`.
5. – Costruzione del grafo finale con l'insieme delle collezioni dei vertici
– e l'insieme delle collezioni degli archi.

This 00:45 | All 07:10 | [Go to next slide](#)

The graph

A subgraph of the obtained final graph:



2021-09-23

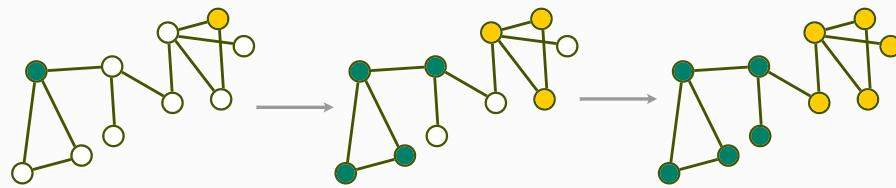
└ The graph



1. Questo è un assaggio del grafo finale, o anche una verifica della sua validità.
2. Il grafo finale è composto dall'insieme di tutti i vertici del dataset e da tutte le connessioni derivanti dalle informazioni dei loro attributi.
3. In questo sottografo sono mostrate le connessioni di primo e secondo grado del professor Previdi, il vertice in mezzo.
4. Si può notare in alto a destra professor Mazzoleni e Scandella.
5. In alto a sinistra professor Ermidoro e in basso a destra professor Bitanti.

This 00:40 | All 07:50 | [Go to next slide](#)

Pregel Label Propagation Community Detection Algorithm



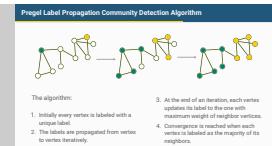
The algorithm:

1. Initially every vertex is labeled with a unique label.
2. The labels are propagated from vertex to vertex iteratively.

3. At the end of an iteration, each vertex updates its label to the one with maximum weight of neighbor vertices.
4. Convergence is reached when each vertex is labeled as the majority of its neighbors.

2021-09-23

Pregel Label Propagation Community Detection



1. Ogni vertice è etichettato con una label unica.
2. Le labels sono fatte propagare da un vertice all'altro in modo iterativo.
3.
 - Alla fine di ogni iterazione di propagazione, un vertice aggiorna la sua label tale che corrisponda a quella col peso massimo dei vertici vicini e loro connessioni.
 - I pareggi sono decisi a random.
4.
 - Si raggiunge un punto di convergenza quando ogni nodo è etichettato come la maggioranza dei suoi vicini.
 - Può succedere che la convergenza ad un'unica soluzione non avvenga in un tempo ragionevole.
 - In tali casi, impostare un numero massimo di iterazioni può essere un trade-off tra accuratezza e tempo d'esecuzione.

This 00:45 | All 08:35 | [go to next slide](#)

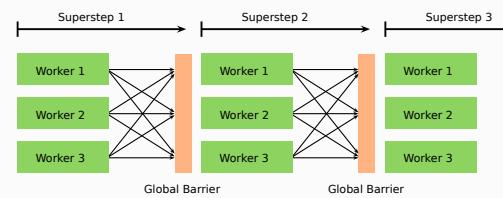
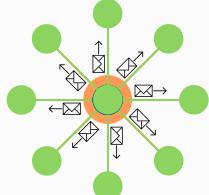
A generic Pregel algorithm

Pregel's supersteps:

Diffusion: information is propagated from vertex to neighbors.

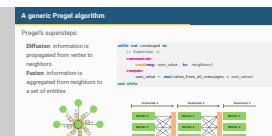
Fusion: information is aggregated from neighbors to a set of entities

```
while not converged do
    /* Superstep */
    communicate:
        send(msg: own_value, to: neighbors)
    compute:
        own_value ← max(value_from_all_messages ∪ own_value)
end while
```



2021-09-23

A generic Pregel algorithm



- Un generico algoritmo Pregel lavora in ambienti distribuiti,
- con computazioni e dati sparsi in diverse macchine.
- Per convergere ad uno stato di consenso distribuito,
- Pregel funziona con iterazioni di superpassi.

I superpassi sono due:

1.
 - La fase di diffusione: l'informazione sulle etichette viene propagata da vertice a vertice.
 - Questo step è anche detto di comunicazione del label.
2.
 - L'altra fase è quella di fusione,
 - o di aggregazione dell'informazione dai vertici ad un insieme di nodi.

This 00:40 | All 9:15 | [Go to next slide](#)

Clustering results

Vertex type	Number of vertices	Number of detected communities
author	2786113	177592
editor	43644	9837
institution	56918	25415
journal	1905	1896
publication	5662747	141939
publisher	2292	1437
school	2098	1677
series	1742	934
all types	8557459	187451

2021-09-23

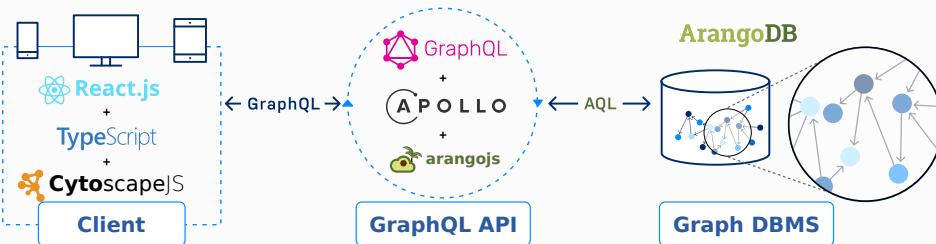
Clustering results

- Nel grafo fatto da 8 milioni e mezzo di vertici e 24 milioni di archi,
 - sono state individuate 187 mila comunità.
- Mediamente in una community ci sono
 - 16 ricercatori, 2 istituti di affiliazione, 1 journal.
 - Inoltre, generalmente i nodi di una community hanno fatto mediamente 40 pubblicazioni scientifiche e le pubblicazioni sono associate alla stessa scuola.
- Per vedere questi risultati in maniera visualmente apprezzabile,
 - è stata sviluppata una Full-stack Web Application.
 - Vediamo come è fatta:

This 00:35 | All 09:50 | [Go to next slide](#)

Clustering results		
Vertex type	Number of vertices	Number of detected communities
author	2786113	177592
editor	43644	9837
institution	56918	25415
journal	1905	1896
publication	5662747	141939
publisher	2292	1437
school	2098	1677
series	1742	934
all types	8557459	187451

Web Application's Architecture



2021-09-23

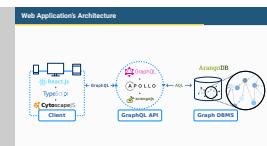
Web Application's Architecture

- La Web Application è composta dall'interfaccia lato Frontend
 - e dall'API in backend che fa uso del graph database.
- Il client è implementato con React, TypeScript e Cytoscape per il rendering dei grafi.
- L'API è sviluppato con GraphQL ed è servito da Apollo Server.
- Arangojs è un driver JavaScript che viene usato per comunicare con ArangoDB.
- Il database è ArangoDB, un DBMS multi model, che si presta da graph database.
- Il linguaggio di querying del database è AQL, ArangoDB Querying Language.
- 3 macchine servono la WebApp, una per il FE, una per il BE e l'ultima per il DB.

Wait 10 seconds

- Fino ad ora abbiamo visto la trasformazione dei dati in grafo con ArangoDB.
- Ora vediamo come è fatto l'API e poi successivamente il frontend.

This 00:50 | All 10:40 | [Go to next slide](#)



GraphQL API



The API, initially REST then converted to a GraphQL API, is made of two resolvers:

1. A resolver function to handle the search form autocomplete suggestions.
2. A resolver function to provide collaboration graph's data.

2021-09-23

└ GraphQL API

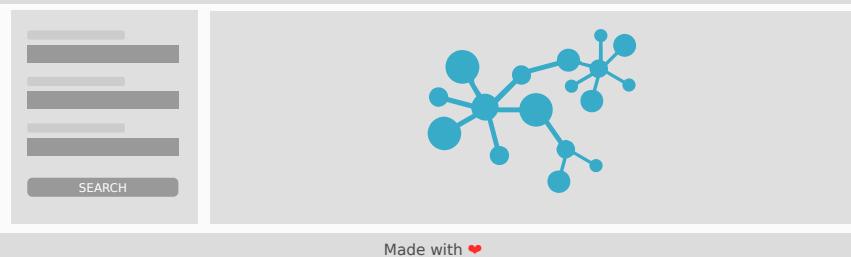
1. – Il lato back-end della Web Application, l'API
 - è stato sviluppato con NodeJS, Express, inizialmente REST API
 - poi convertito in un GraphQL API con Apollo Server.
2. – GraphQL è un linguaggio di querying e manipolazione dati per APIs.
 - Fa uso di un sistema a tipi e campi per gestione delle richieste.
3. – L'API implementato è composto da due resolver functions.
 - Uno per la fornire i suggerimenti di autocompletamento durante la ricerca.
 - L'altro per fornire i dati (vertici, archi e communities)
 - che compongono il grafo delle collaborazioni.

This 00:35 | All 11:15 | [Go to next slide](#)



Frontend UI Layout

Academic Graph Connections

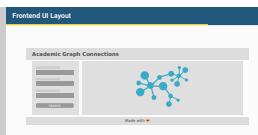


2021-09-23

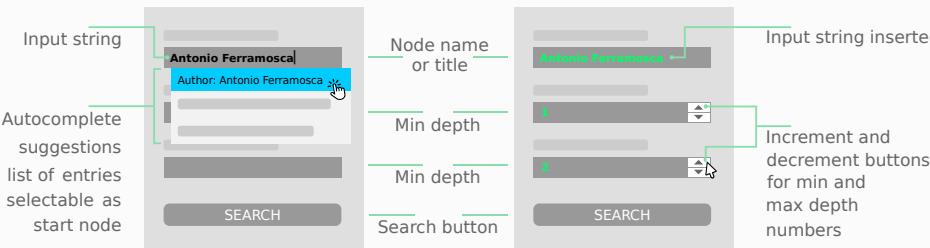
└ Frontend UI Layout

1. – La client interface, stilata mediante bootstrap, ha un header,
 - un content principale suddiviso in due colonne e un footer.
2. – Nella colonna a sinistra del content, l'utente può inserire i dati
 - per ricercare il grafo delle comunità di collaborazione di un nodo.
3. – Nella colonna a destra invece, una volta interrogato l'API
 - e ricevuta la risposta, viene renderizzato il grafo ricercato.
4. Vediamo più in dettaglio il form di ricerca.

This 00:35 | All 11:50 | [Go to next slide](#)



Frontend UI Search Form



2021-09-23

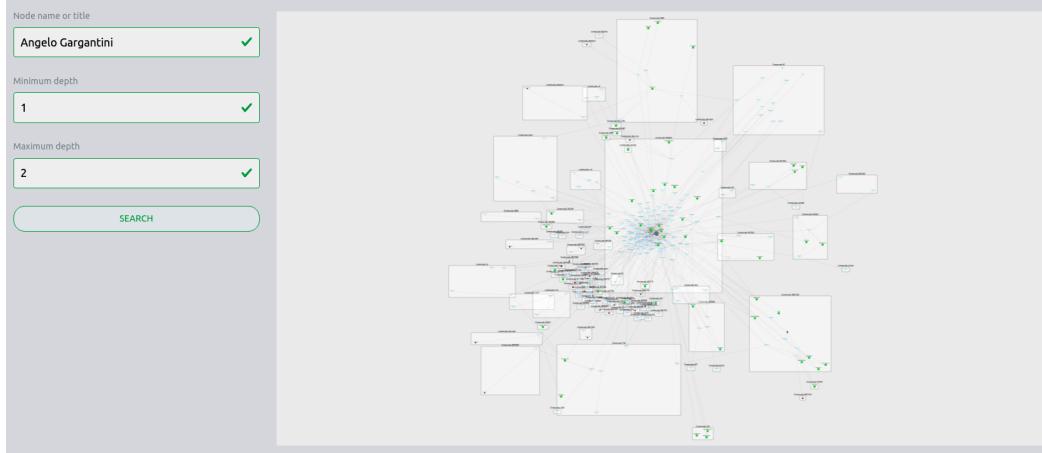
Frontend UI Search Form

1.
 - La ricerca è implementata con una feature di autocompleteamento
 - in modo che l'utente possa selezionare un nodo dall'elenco dei nodi suggeriti
 - e in background il suo nome o titolo verrà tradotto nel suo ID.
 - Tale ID poi viene usato durante la richiesta del grafo delle collaborazioni.
2.
 - Nell'esempio mostrato in slide
 - si è cercato il nodo rappresentante un autore di pubblicazioni scientifiche
 - di nome "Antonio Ferramosca" *smile*
3.
 - Gli altri due campi da compilare sono il minimum e il maximum depth.
 - Essi rappresentano le distanze minime e massime che i nodi da visualizzare devono avere dal nodo di partenza.
4.
 - Vediamo un caso concreto.
 - Ricerchiamo il grafo delle comunità di collaborazione del professor Gargantini.

This 00:55 | All 12:45 | [Go to next slide](#)

Results display (Searching for "Angelo Gargantini")

Academic Graph Connections



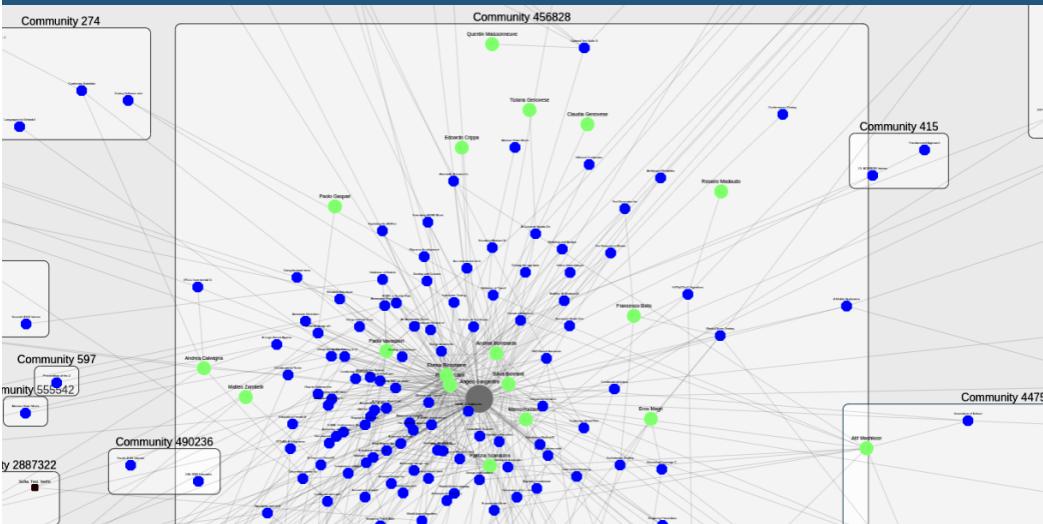
2021-09-23

Results display (Searching for "Angelo Gargantini")

1. Nella slide è mostrata l'interfaccia che l'utente vede mentre ricerca il grafo delle communities di collaborazione del professor Gargantini.
2. A sinistra c'è la sezione dedicata alla ricerca di cui si è parlato nella slide precedente.
3.
 - A destra invece, una volta finito il querying dell'API
 - e il fitting dei node e archi del grafo,
 - viene visualizzato il grafo delle comunità di collaborazione.
4.
 - È facile notare i rettangoli che rappresentano le communities.
 - Dentro un rettangolo sono i vertici membri di quella community.
 - Zoomando in mezzo si può vedere ...

This 00:40 | All 13:25 | [Go to next slide](#)

Results display (Angelo Gargantini)

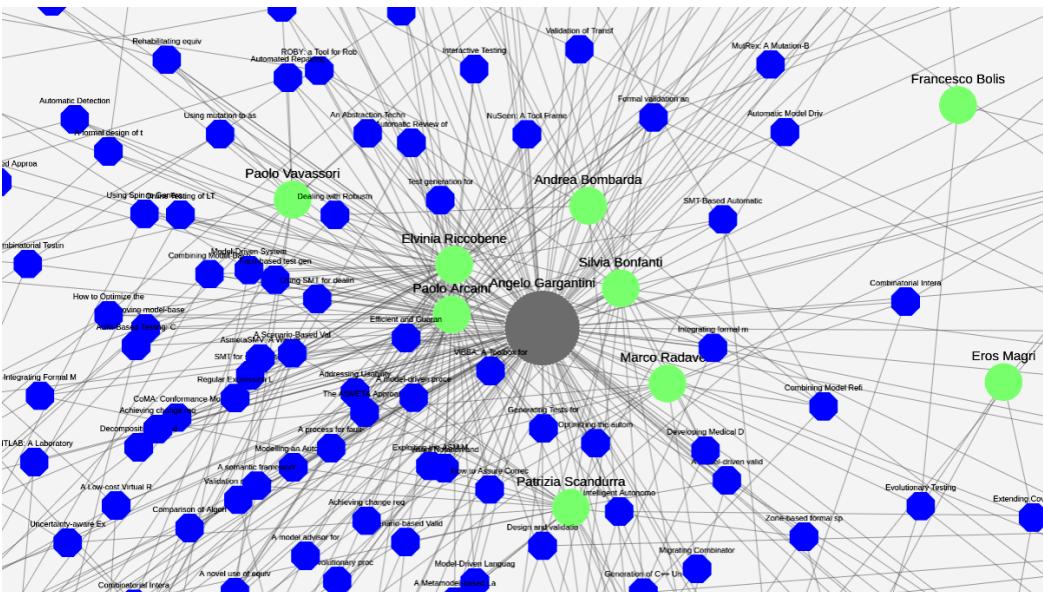
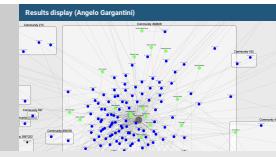


2021-09-23

└ Results display (Angelo Gargantini)

1. ... come ogni community sia contraddistinta da un ID numerico.
 2. I nodi verdi sono autori, quelli blu sono le pubblicazioni scientifiche.
 3.
 - Un aspetto che può far sorgere dei dubbi
 - sono le comunità mostrate come composte da uno o due nodi.
 - In realtà esse hanno più nodi, ma perché gli altri nodi distanti dal nodo ricercato, più di un arco, un salto
 - allora essi non vengono inclusi.
 - In ogni caso, comunità composte da pochi nodi
 - fanno comunque senso dal punto di vista dell'algoritmo di community detection.

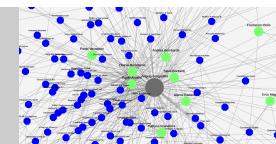
This 00:35 | All 14:00 | Go to next slide



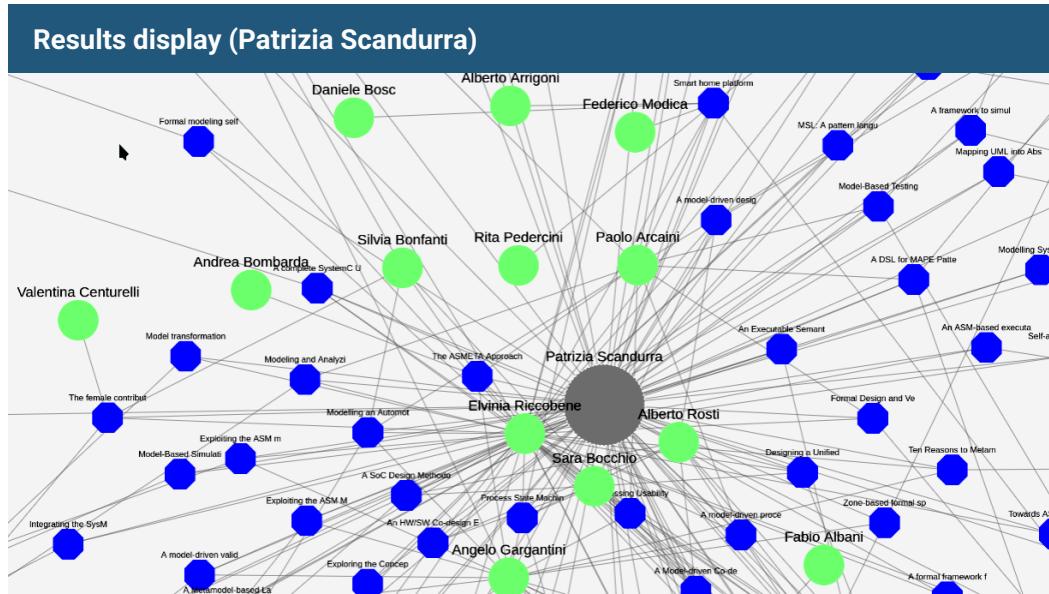
2021-09-23

1. – Zoomando ancora, dentro il rettangolo rappresentante la collaboration community,
 - si possono leggere i nomi dei ricercatori con cui professor Gargantini collabora
 - e anche i titoli delle pubblicazioni appartenenti a questo cluster.
 2. – Uno dei nodi parte del grafo della comunità di collaborazione scientifica del professor Gargantini
 - è anche la professoressa Scandurra.
 - Se andassimo a ricercare la community della professoressa ...

This 00:35 | All 14:35 | Go to next slide



Results display (Patrizia Scandurra)

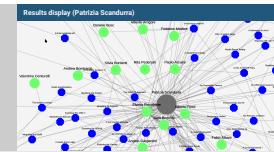


2021-09-23

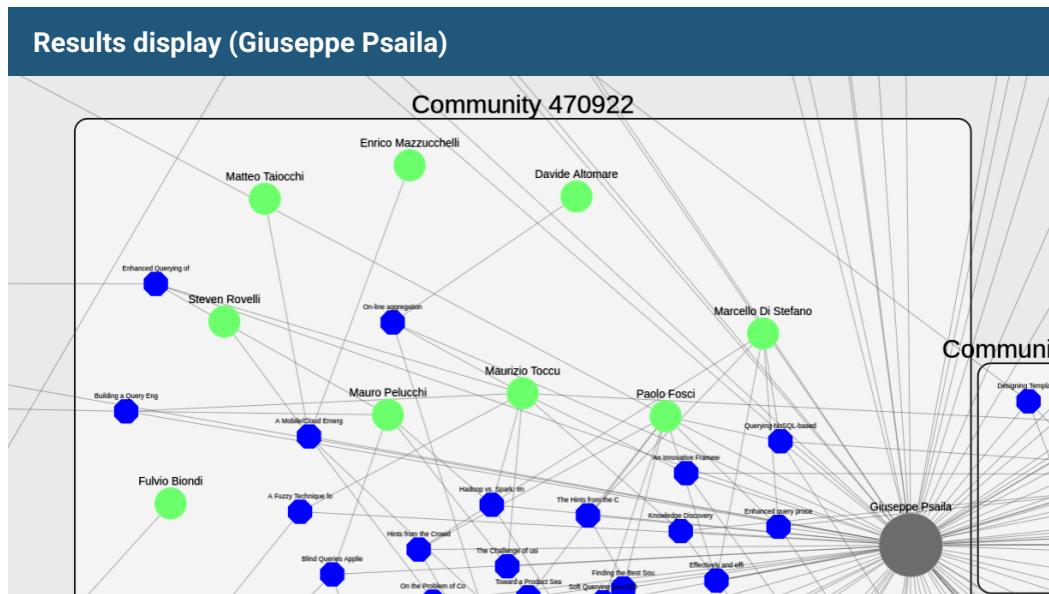
Results display (Patrizia Scandurra)

- ... sarebbe possibile vedere che la community è la stessa.
- Nonostante il posizionamento, il fitting dei nodi sia diverso,
- la community è costituita degli stessi nodi e archi.
- È interessante far notare che essere docenti nella stessa facoltà e
stesso dipartimento,
- non necessariamente si traduce in appartenenza alla stessa
comunità di collaborazione scientifica.
- Ad esempio, cercando il grafo di collaborazione del professor Psaila ...

This 00:35 | All 15:10 | Go to next slide



Results display (Giuseppe Psaila)

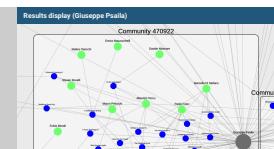


2021-09-23

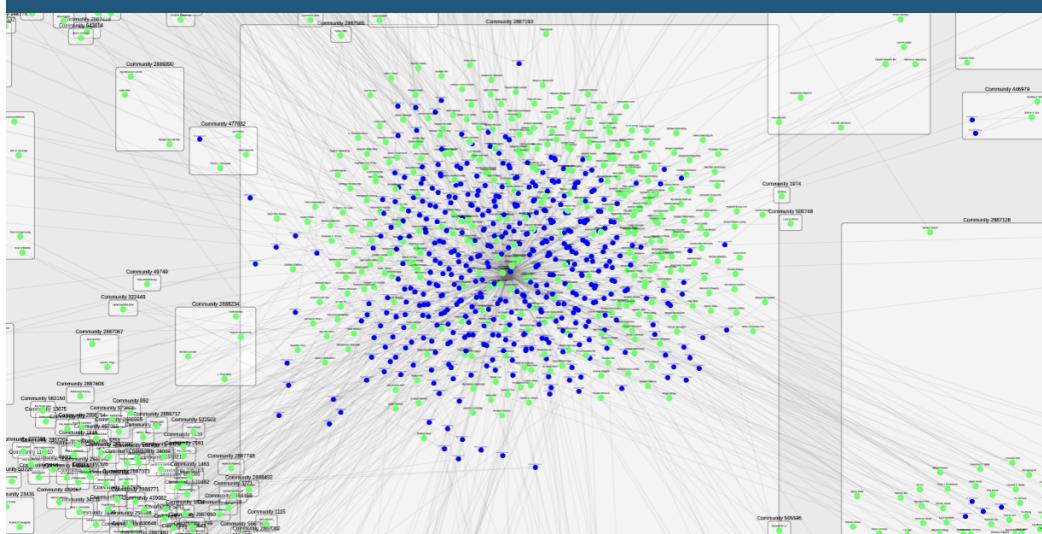
Results display (Giuseppe Psaila)

- ... è possibile riconoscere subito i suoi assistenti, Fosci e Pelucchi
- ma non sono presenti i docenti visti poc'anzo,
- ovvero professor Gargantini o professoressa Scandurra.
- Questo succede perché nei dati, di fatto le relazioni (gli archi)
- tra loro e il nodo di professor Psaila sono relativamente più sparsi,
- meno densi di quanto lo siano con altri ricercatori.
- Dal LPA essi quindi vengono individuati, correttamente,
- come parte di due comunità di collaborazione scientifica distinte.
- Nella successiva slide verrà mostrato il grafo delle comunità di
collaborazione accademica ora non più di un autore specifico ...

This 00:40 | All 15:50 | Go to next slide



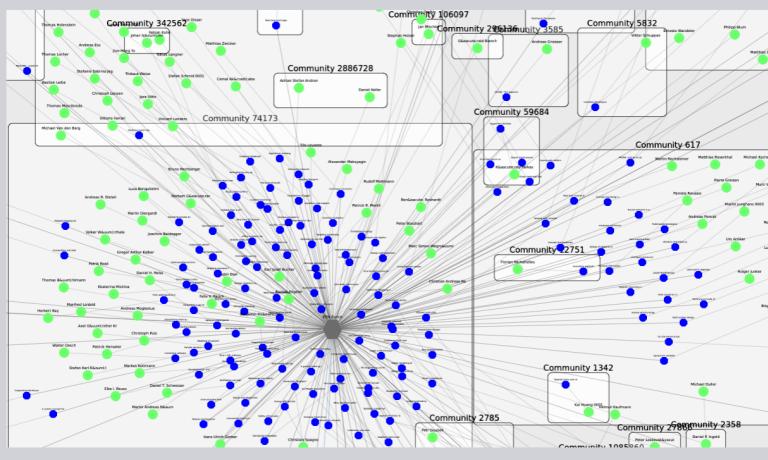
Results display (Statistical Methods & Applications Journal)



Results display (ETH Zurich)

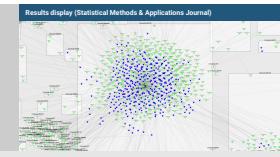
Academic Graph Connections

Node name or title	ETH Zurich	✓
Minimum depth	1	✓
Maximum depth	2	✓



30031-00-23

↳ Results display (Statistical Methods & Applications)



1. ... ma di un Journal come quello di Statistical Methods & Applications.
 2. Uno dei nodi verdi, se zoomassimo, è professor Fassò.
 3. Per concludere ...

This 00:15 | All 16:05 | Go to next slide

30031-00-23

└ Results display (ETH Zurich)



- ... a scopo illustrativo della completezza del lavoro svolto,
 - viene mostra in questa slide come non solo si possano cercare comunità di collaborazione scientifica costruite attorno a ricercatori
 - o journals, o editors - ma anche communities legate ad una scuola o un istituto accademico di affiliazione.
 - Nel caso specifico sono mostrate le comunità di collaborazione scientifica della più prestigiosa università di informatica in Europa,
 - ovvero l'ETH di Zurigo.
 3. Con questa slide finisce la presentazione...

This 00:35 | All 16:40 | Go to next slide

Name
Surname

**CLUSTERING
GRAPHS**

Applying a Label Propagation
Algorithm to Detect Communities
in Graph Databases

Thank you!

Questions?

2021-09-23

1. Grazie mille a tutti per l'attenzione!
2. Se avete delle domande...

This 00:20 | All 17:00 | wait for questions

Thank you!

Questions?

Name

Surname

**CLUSTERING
GRAPHS**

Applying a Label Propagation
Algorithm to Detect Communities
in Graph Databases