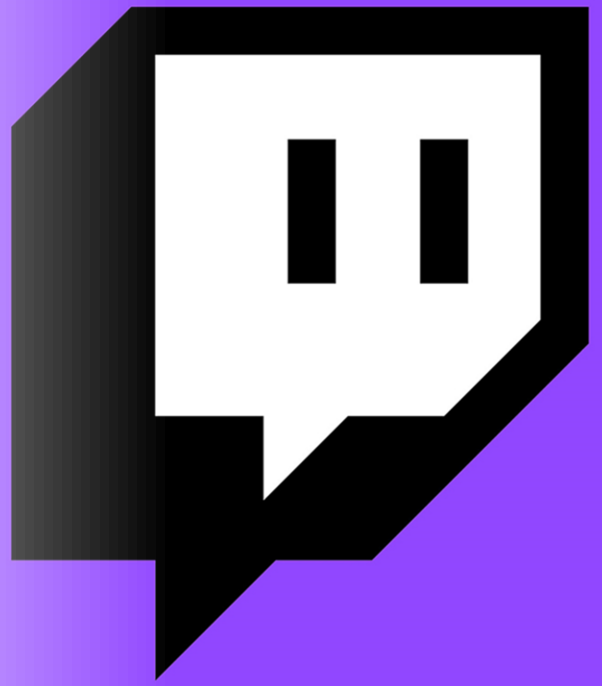
Università degli Studi di Milano – Bicocca
Corso di Laurea Magistrale in Data Science
Anno Accademico 2021/2022

Progetto di Data Management

ANALISI DELLE COMMUNITY ITALIANE DI TWITCH



Progetto di:
Silvia GROSSO matricola n. 881993
Paola IMPICCICHÈ matricola n. 878163
Gianluca SCURI matricola n. 886725

Introduzione a Twitch

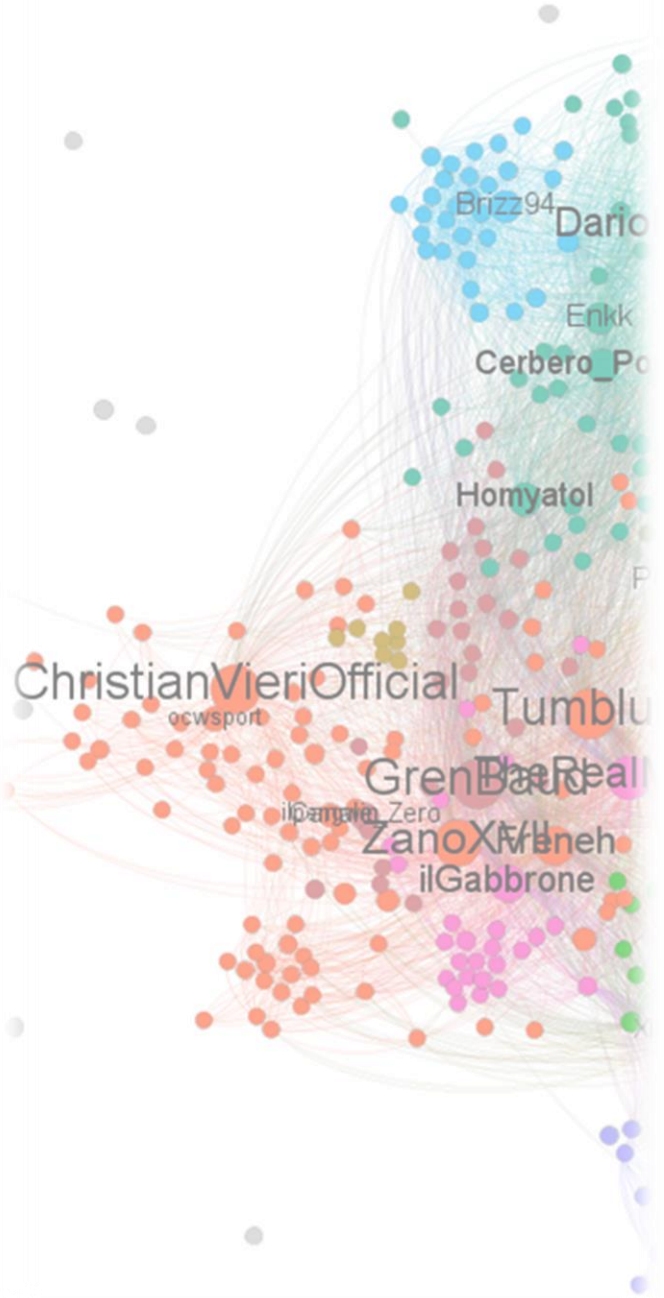
- Lanciato nel **2011** da Justin Kan e Emmett Shear ed acquistato nel **2014** da **Amazon** per **970 milioni** di **dollari**
- In **Italia** sono circa **3 milioni** gli utenti che fruiscono regolarmente dei contenuti
- Possibilità di **guadagno** per gli **streamer** tramite abbonamenti, donazioni, pubblicità
- Piattaforma come **vetrina di sponsorizzazione** di prodotti per le aziende

eSports **videogiochi**
musica
IRL creatività

Obiettivi

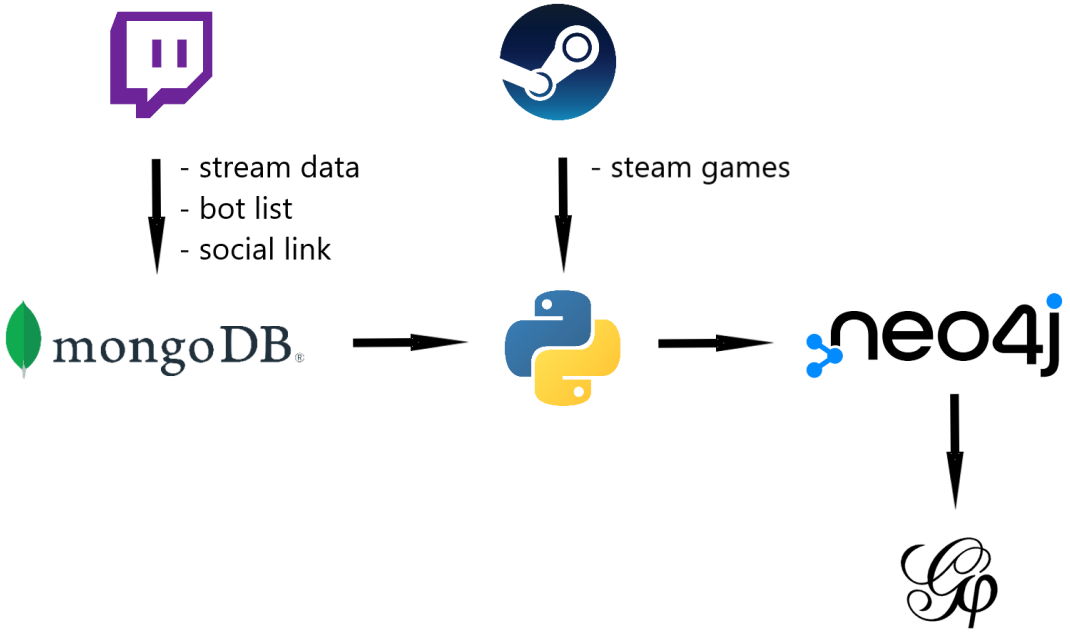
- Realizzazione di un **modello a grafo** utile per
  - Streamer neofiti
  - Streamer affermati
  - Inserzionisti pubblicitari

- Intenti di ricerca:
  - Analisi e visualizzazione dei **cluster**
  - Analisi della **qualità di interazione** tra lo streamer e gli utenti spettatori
  - Analisi e visualizzazione della **concorrenza** tra community
  - Analisi della **reattività** di una community alla **pubblicazione** di un **nuovo gioco**



Struttura del progetto

- Acquisizione dei dati da **Twitch** in tempo reale
- Raccolta dei dati in formato documentale su **MongoDB**
- Acquisizione dei dati da **SteamDB**
- Exploration, Processing e Modeling in **Python**
- Storage su **Neo4J**
- Query e visualizzazione su **Gephi** del database a grafo ottenuto



Data Acquisition - Twitch

- **Stream Data**

`[streamer, spect, game_name, viewer_count]`

  - API REST ufficiali
  - Dal 05/05/2022 al 19/05/2022 ogni 5 minuti
  - 3823 istanti raccolti (su MongoDB)

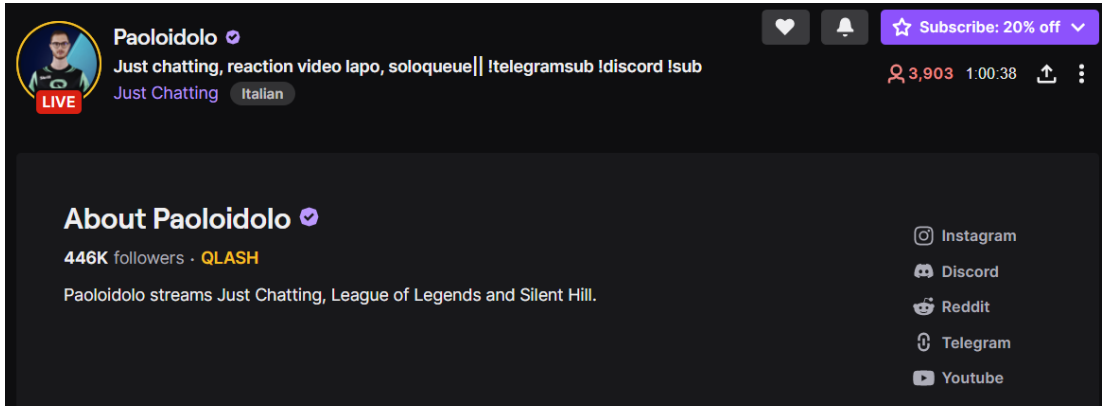
- **Link social**

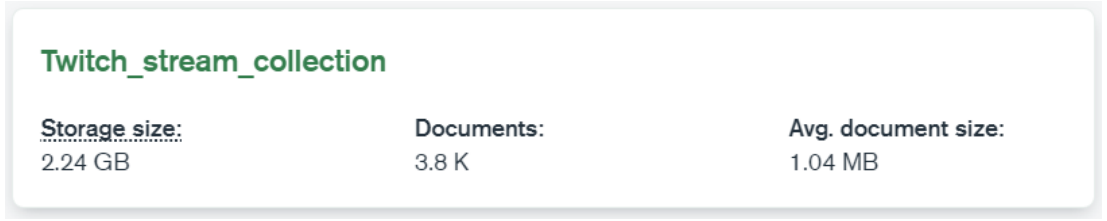
`[streamer, social_list]`

  - API REST
  - 2945 streamer

- **Lista Bot**

  - Scraping con estensione web
  - 5576 nickname



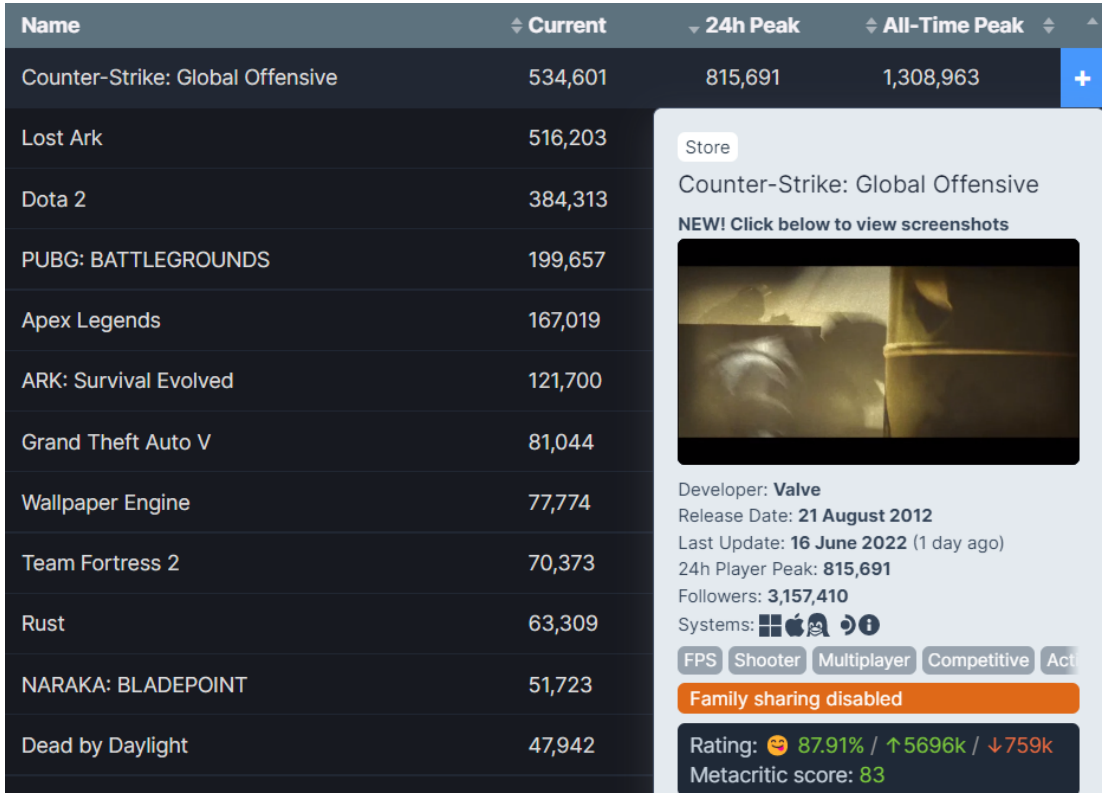


Data Acquisition - SteamDB

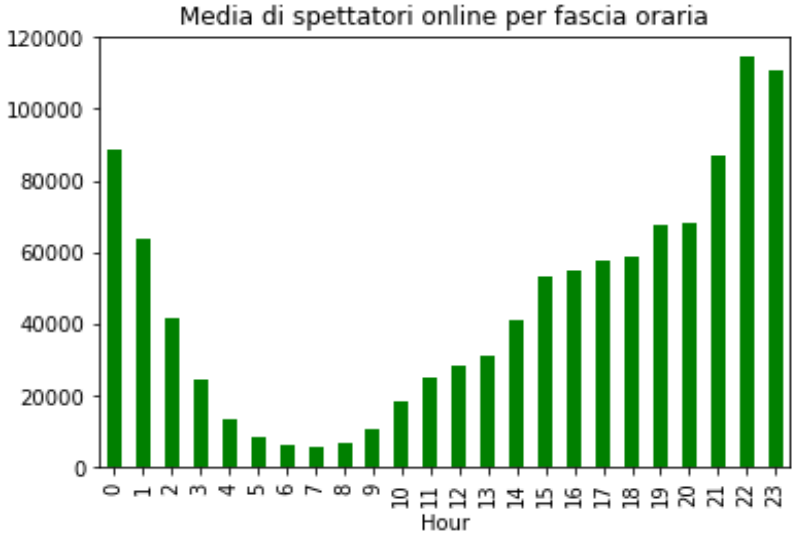
- **Games dataset**

`[Title, Developer, Publisher, Release Date, 24h Player Peak, Followers, Categories]`

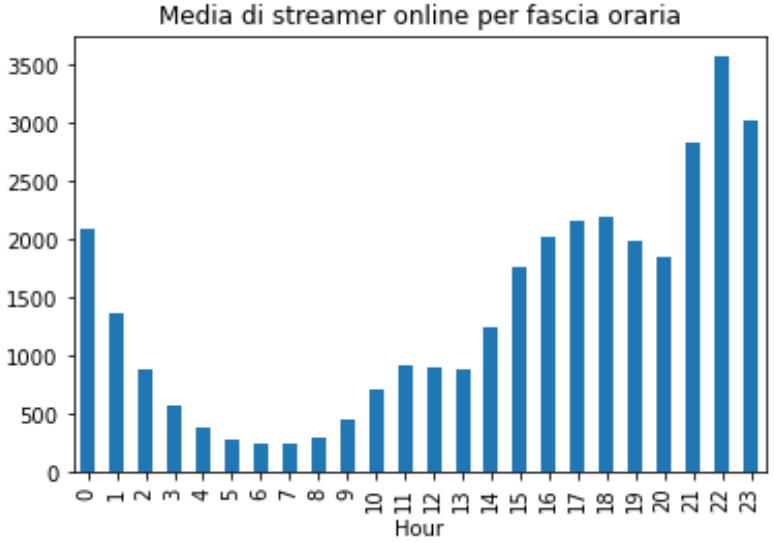
- Scraping dinamico con Selenium della pagina “Most Played Games”
- 3498 giochi

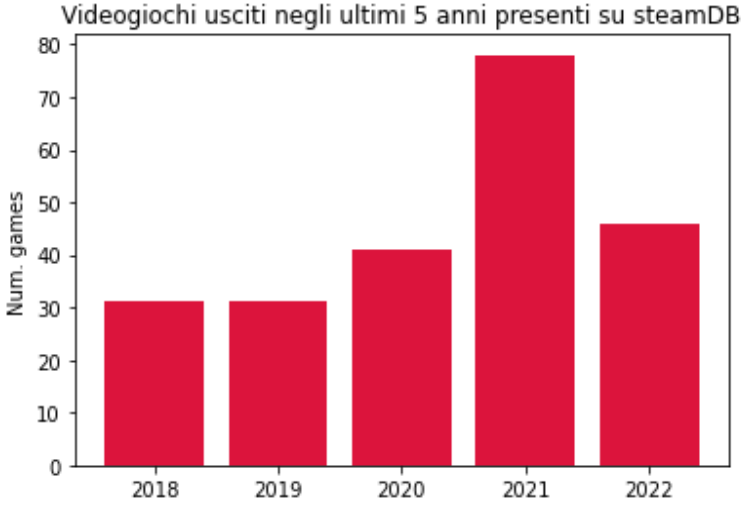


Data Exploration



	minuti trascorsi in live	spettatori totali
mean	1466.684582	1656.022506
std	1735.805536	7450.436019
min	15.000000	10.000000
25%	345.000000	82.000000
50%	960.000000	231.000000
75%	2010.000000	801.000000
max	19125.000000	157718.000000





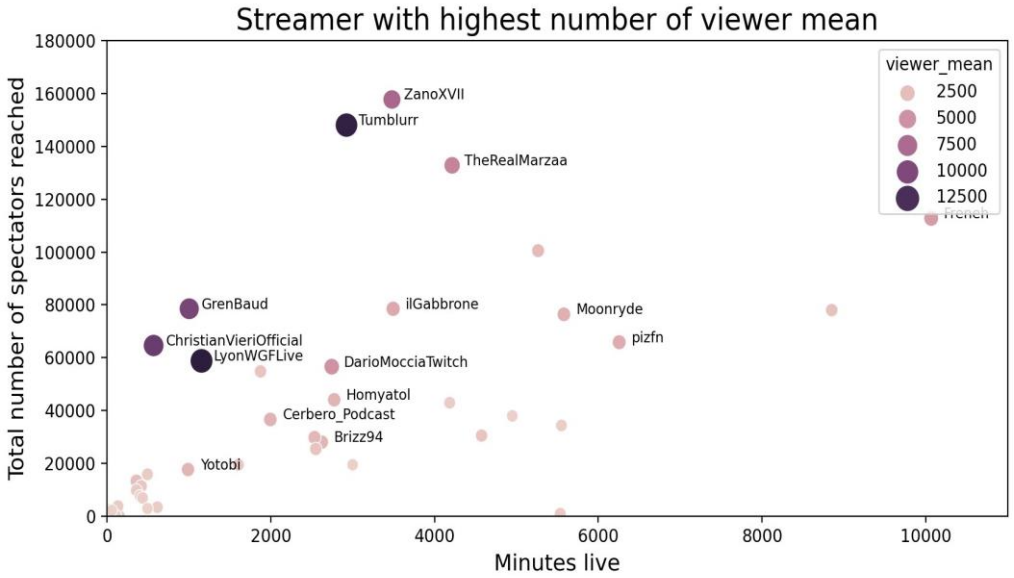
Data Processing

- Rimozione bot
- Espansione e aggregazione (da istanti a streamer)
- Calcolo lista categorie, lista spettatori e statistiche (valori medi e picchi massimi di viewer e spettatori)
- Creazione **Dataset Streamer** e **Dataset games** con assegnazione chiave univoca
- Creazione **Dataset Streamer-Games** calcolando minuti di live relativi
- Creazione **Dataset Streamer-Streamer** calcolando overlap percentuale a partire dagli spettatori comuni

$$overlap = \frac{num\ spect\ a}{num\ spect\ b} \times 100, con\ a < b$$

- Esportazione di tutti i dataset in formato CSV

1. **Dataset Streamer**
   - `idStreamer`
   - `streamer`
   - `minutesLive`
   - `viewerMean`
   - `viewerPeak`
   - `spectMean`
   - `spectTot`
2. **Dataset Games**
   - `idGame`
   - `gameName`
3. **Dataset Streamer-Games**
   - `ID_streamer`
   - `ID_game`
   - `minutes`
4. **Dataset Streamer-Streamer**
   - `ID_streamer_i`
   - `ID_streamer_j`
   - `overlap_percentage`



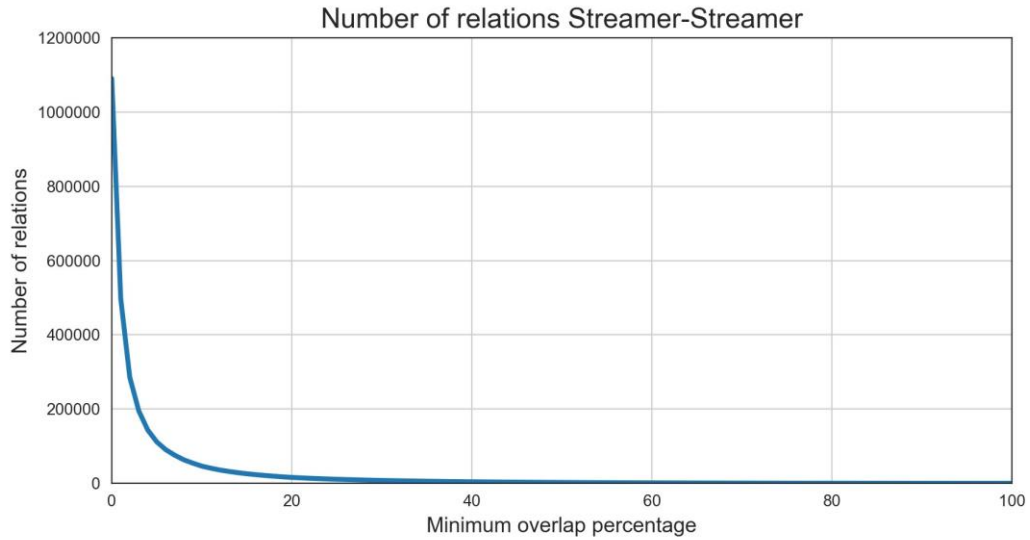
Data Processing

- **Soglie** minime di attività per ogni **viewer** :
  - 30 minuti di visione della live
- Soglie minime di attività per ogni **streamer** :
  - 60 minuti di live nel periodo
  - 10 spettatori medi
  - 10 viewer medi
  - 30 minuti di live per categoria
  - 10% di overlap con un altro streamer

45583 streamer → **2977** streamer

1102634 relazioni → **47222** relazioni

```
interval = 5                        # Interval between observations (fixed)
freq = 3                            # Sample frequency (1=5min, 2=10min, 3=15min ecc.)
min_stream_time = 4                 # Minimum live time parameter
min_watch_time = 2                  # Minimum threshold of time as a spectator (<= min_stream_time)
min_viewer_mean = 10                # Minimum number of mean viewers
min_spect_mean = min_viewer_mean    # Minimum number of mean spectator
min_overlap_percentage = 10         # Minimum overlap percentage between 2 streamers
min_game_time = 2                   # Minimum number of minutes streamed for each category/game
```



Data Enrichment

- **Link Social Network** :

Left outer join tra i **2977** streamer in **Streamer_dataset** e i **2585** nel dataset **social_link**

Streamer

<id>	5275
Name	Homyatol
idStreamer	12
minutesLive	2775
socialLinks	https://www.instagram.com/homywankenobi/?hl=it
spectMean	2409
spectTot	44082
viewerMean	3060
viewerPeak	6409

- **Informazioni di SteamDB** :

Left outer join tra **1144** categorie di **Games_dataset** e **3498** giochi contenuti nel file JSON **steam_games**

Game

<id>	2301
Name	Stranded Deep
categories	Open World Survival Craft; Survival; Multiplayer; Open World; Crafting; First-Person
developer	Beam Team Games
idGame	15
publisher	Beam Team Publishing
releaseDate	2015-01-23
steamName	Stranded Deep

Data Enrichment

twitch_name	steam_name	twitch_clean	steam_clean	match_score
Monkey Island 2 Special Edition: LeChuck's Rev...	Monkey Island™ 2 Special Edition: LeChuck's Re...	monkey island 2 special edition lechuck's revenge	monkey island 2 special edition lechucks revenge	100
Assassin's Creed IV: Black Flag	Assassin's Creed® IV Black Flag™	assassin's creed iv black flag	assassins creed iv black flag	100
BlazBlue: Central Fiction	BlazBlue Centralfiction	blazblue central fiction	blazblue centralfiction	100
Tom Clancy's The Division	Tom Clancy's The Division™	tom clancy's the division	tom clancys the division	100
Assassin's Creed: Brotherhood	Assassin's Creed® Brotherhood	assassin's creed brotherhood	assassins creed brotherhood	100

- Esempi di coppie dello stesso gioco con differenze sintattiche nelle stringhe

twitch_name	steam_name
Assassin's Creed IV: Black Flag	Assassin's Creed® IV Black Flag™
BlazBlue: Central Fiction	BlazBlue Centralfiction
Tom Clancy's The Division	Tom Clancy's The Division™
Assassin's Creed: Brotherhood	Assassin's Creed® Brotherhood
Puyo Puyo Tetris	Puyo Puyo™Tetris®
MotoGP 22	MotoGP™22
Stronghold Crusader II	Stronghold Crusader 2
Dark Messiah of Might and Magic	Dark Messiah of Might & Magic

- Esempi di coppie di giochi differenti con stringhe sintatticamente simili

twitch_name	steam_name
Infernax	Inferna
Spellbreaker	Spellbreak
Cyber Hunter	Cyberhunt

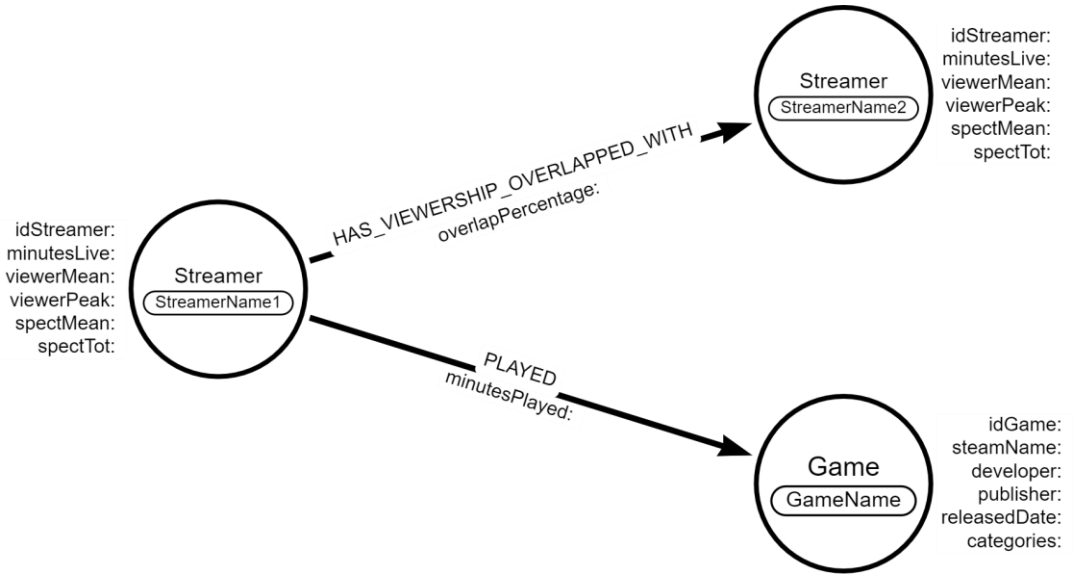
Data Modeling and Storage - Struttura

- **Creazione constraint**

```
CREATE CONSTRAINT StreamerNameKey IF NOT EXISTS
FOR (s:Streamer)
REQUIRE s.Name IS UNIQUE
```

- **Creazione grafo**

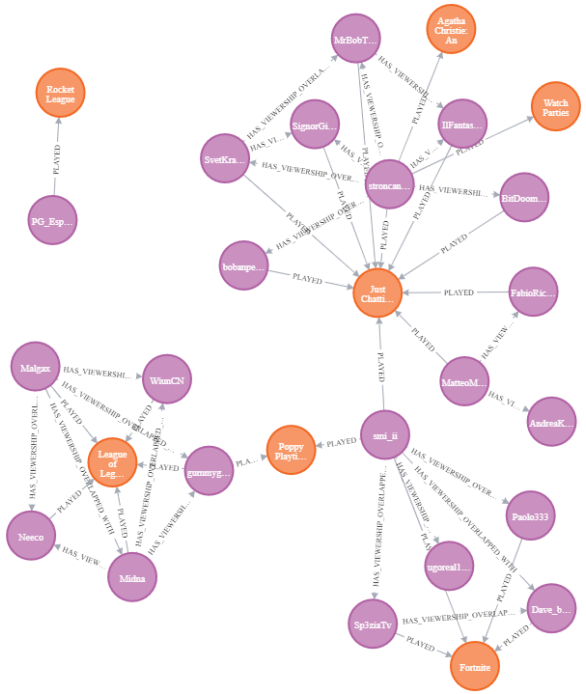
```
LOAD CSV WITH HEADERS FROM "link" AS row
MATCH (i:Streamer {ID_streamer:
row.ID_streamer_i}), (j:Streamer {ID_streamer:
row.ID_streamer_j})
CREATE (i)-[:HAS_VIEWERSHIP_OVERLAPPED_WITH {
overlapPercentage: toFloat(row.
overlap_percentage)}]->(j)
```



Data Modeling and Storage - Query

- Query 1

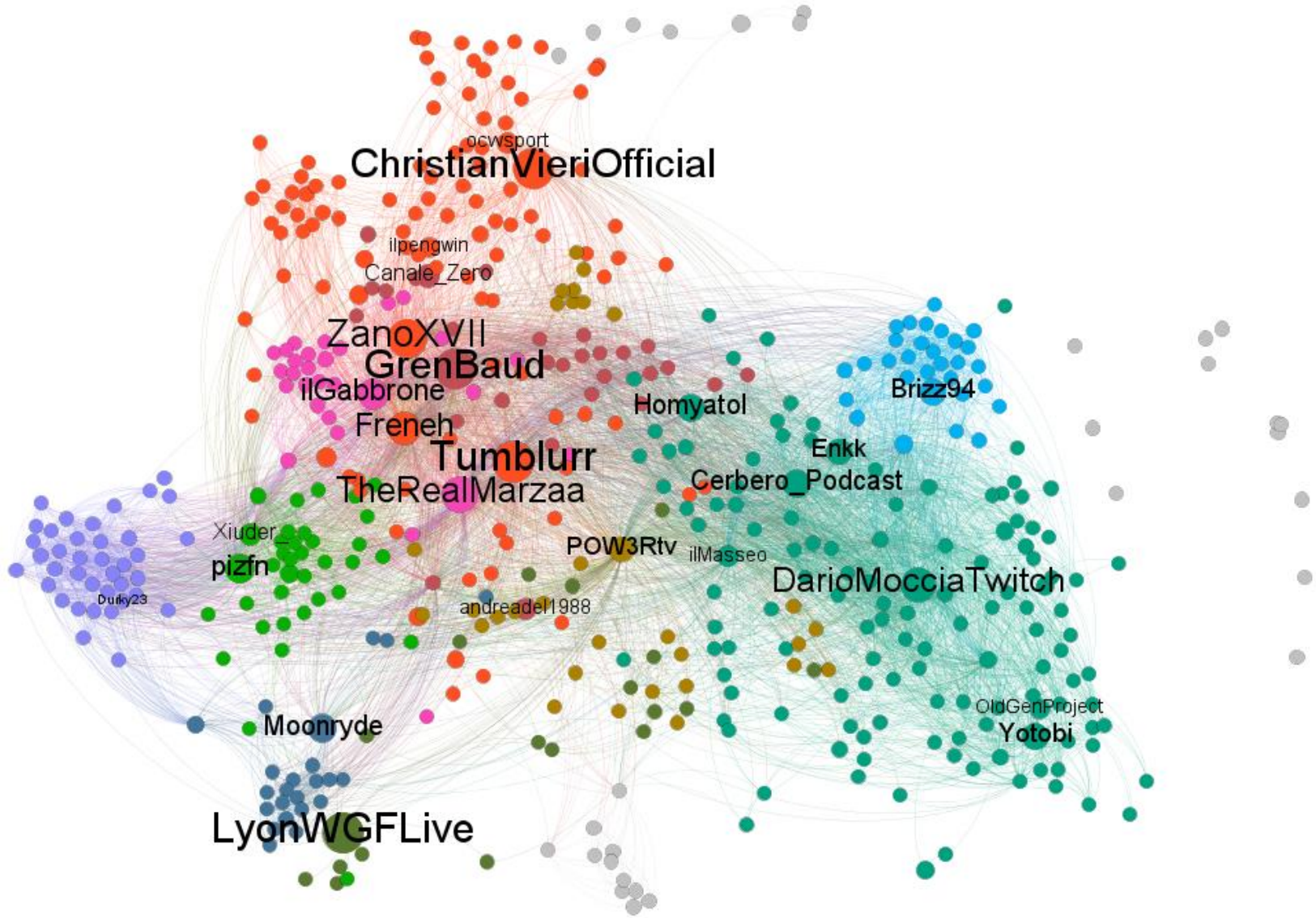
```
MATCH p=()-->()
RETURN p LIMIT 25
```



- Query 2

```
MATCH p=(s:Streamer {Name: 'Homyatol'})-
[r: HAS_VIEWERSHIP_OVERLAPPED_WITH]-()
WHERE r.overlapPercentage > 50
RETURN p
```

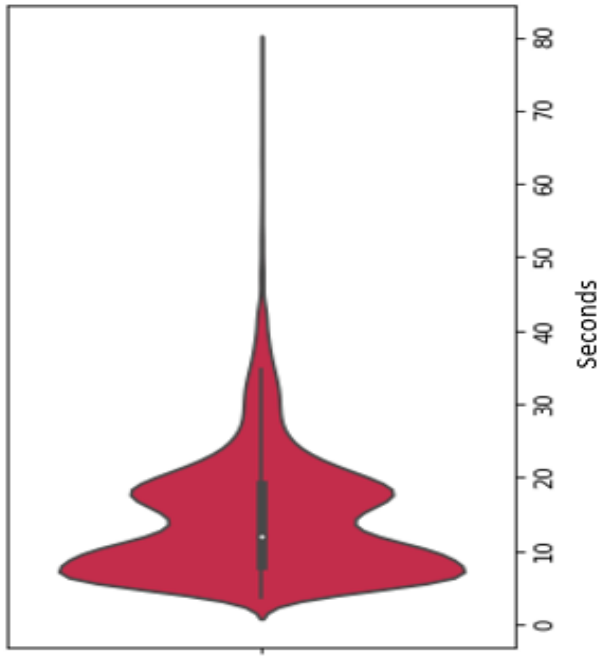
Visualizzazione in Gephi



Data Quality

❑ CURRENCY

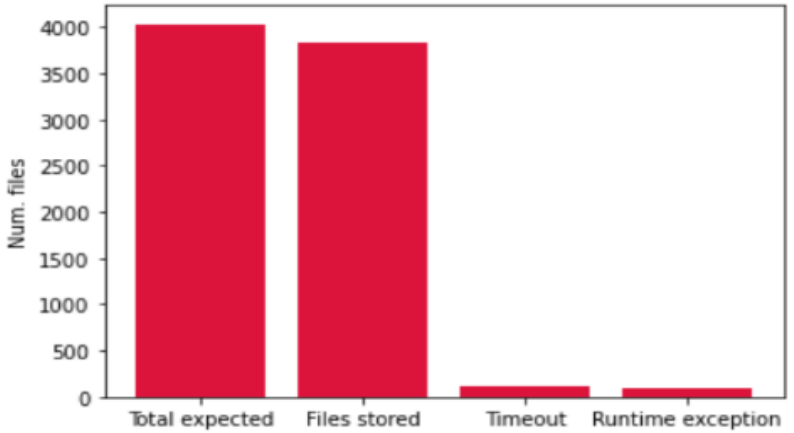
- Differenza tra timestamp di salvataggio del file e timestamp di update



❑ CONSISTENCY

❑ COMPLETENESS

- 5% dei file perso per errori di Timeout o Runtime exception



- Attributi con valori di completeness critici:

Games Dataset	Streamer Dataset
▪ steamName 42% ▪ developer 42% ▪ categories 42% ▪ releaseDate 36% ▪ publisher 20%	▪ social link 87%

Conclusioni

- Il database ottenuto permette di avere un quadro completo di tutto il **panorama italiano di Twitch** dal **05/05/2022** al **19/05/2022**
- La modellizzazione a grafo dei dati ha permesso di evidenziare le **relazioni tra le entità**
- Il database può essere **interrogato molto velocemente** grazie all'indicizzazione dei nodi e alla gestione del numero di relazioni
- È possibile **rieffettuare l'analisi** di un determinato arco temporale semplicemente rieseguendo il codice, eventualmente con parametri e soglie differenti
- Attraverso il database è possibile rispondere a domande come:
  - Quali streamer presentano più **spettatori in comune** rispetto ad un altro streamer?
  - Per quanto tempo si è discusso di una particolare **categoria di gioco** nelle live?
  - La **community** di streamer che tratta strategy game è più **ampia** di quella che tratta simulation game?
- Il risultato permette molteplici applicazioni e sviluppi futuri

Possibili sviluppi futuri

- Utilizzo di **altre fonti** per l'arricchimento delle informazioni sui videogiochi (es. Origin, Battle.net) e ampliamento a **categorie IRL** (ad oggi necessaria text mining o analisi video)
- Enrichment con **eventi**, **competizioni** e **nuove release** riguardo ai videogiochi
- **Sentiment analysis** sulle chat (possibile identificazione community tossiche)
- **Analisi temporale** della popolarità di uno streamer o di un topic:
  - Trend prediction
  - Viewer count prediction
  - Lifetime estimation

Riferimenti

- "Twitch.tv." [Online]. Available: https://twitch.tv/
- "Steamdb." [Online]. Available: https://steamdb.info/graph/
- "Neo4j community edition 4.4.8." [Online]. Available: https://neo4j.com/
- "Mongodb community server 5.0.9." [Online]. Available: https://www.mongodb.com/
- "Gephi 0.9." [Online]. Available: https://gephi.org/
- "Twitch developers." [Online]. Available: https://dev.twitch.tv/
- "Asyncio 3.4.3." [Online]. Available: https://docs.python.org/3/library/asyncio.html
- "Pymongo 4.1.1." [Online]. Available: https://pymongo.readthedocs.io/en/stable/
- "Twitch gql." [Online]. Available: https://gql.twitch.tv/gql
- "Steam." [Online]. Available: https://store.steampowered.com/
- "Selenium library 4.2." [Online]. Available: https://selenium-python.readthedocs.io/
- "Beautifulsoup library 4.9.0." [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Riferimenti

- "Twitch insights." [Online]. Available: https://twitchinsights.net/bots
- "Table capture (estensione chrome browser)." [Online]. Available: https://chrome.google.com/webstore/detail/table-capture/iebpjdmgckacbodjpijphcplhebcmeop
- "Forum viewer and spect count." [Online]. Available: https://help.twitch.tv/s/article/understanding-viewer-count-vs-users-in-chat
- "Pandas 1.4.2." [Online]. Available: https://pandas.pydata.org/
- "Numpy 1.22." [Online]. Available: https://numpy.org//
- "Thefuzz 0.19.0, pypi." [Online]. Available: https://pypi.org/project/thefuzz/
- "Origin." [Online]. Available: https://www.origin.com/ita/en-us/store
- "Battle.net." [Online]. Available: https://eu.shop.battle.net/en-us