Dipartimento di Informatica
Università degli studi di Bari

COMMUNITY DETECTION IN TWEETS

Petruzzelli Alessandro

Esame di Gestione e Anlisi di Big Data
Prof. Gianvito Pio

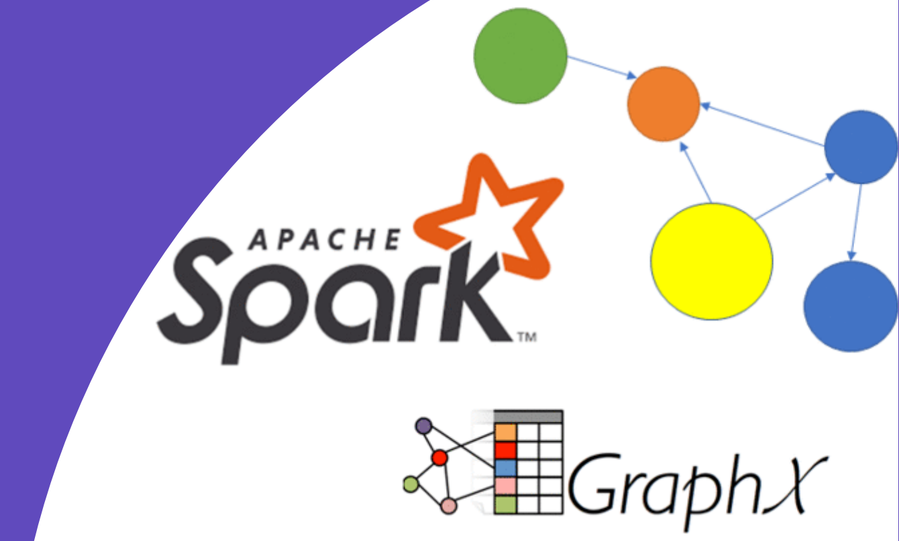
CONTESTO

- I social network hanno portato a raccogliere grandi quantità di dati.
- Le interazioni tra gli utenti sono memorizzate attraverso grafi.
- L'analisi di questi grafi dalle grandi dimensioni sono una sfida perché, spesso, i dati sono salvati in maniera distribuita.

RISORSE

Apache Spark: framework per calcolo distribuito

GraphX: componenteintegrata in Spark che permette l'analisi e il processing di grandi grafi salvati in maniera distribuita



OBIETTIVO

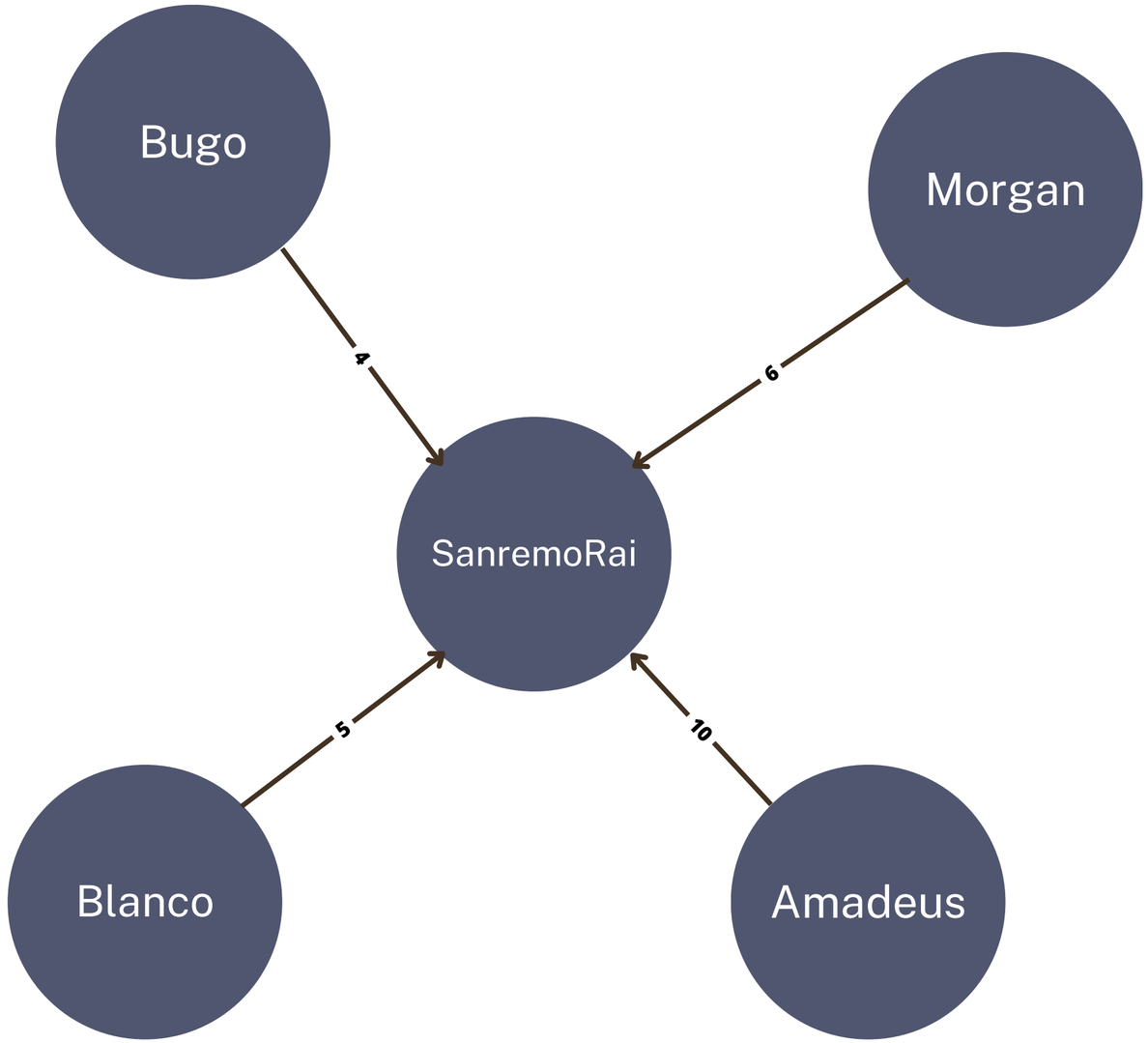
Implementare in maniera distribuita algoritmi di community detection proposti in lettaratura

Gli algoritmi sono stati testati su un dataset di interazioni avvenute sul social network Twitter

Moertini, Veronica S., and Mariskha T. Adithia. "Uncovering Active Communities from Directed Graphs on Distributed Spark Frameworks, Case Study: Twitter Data."

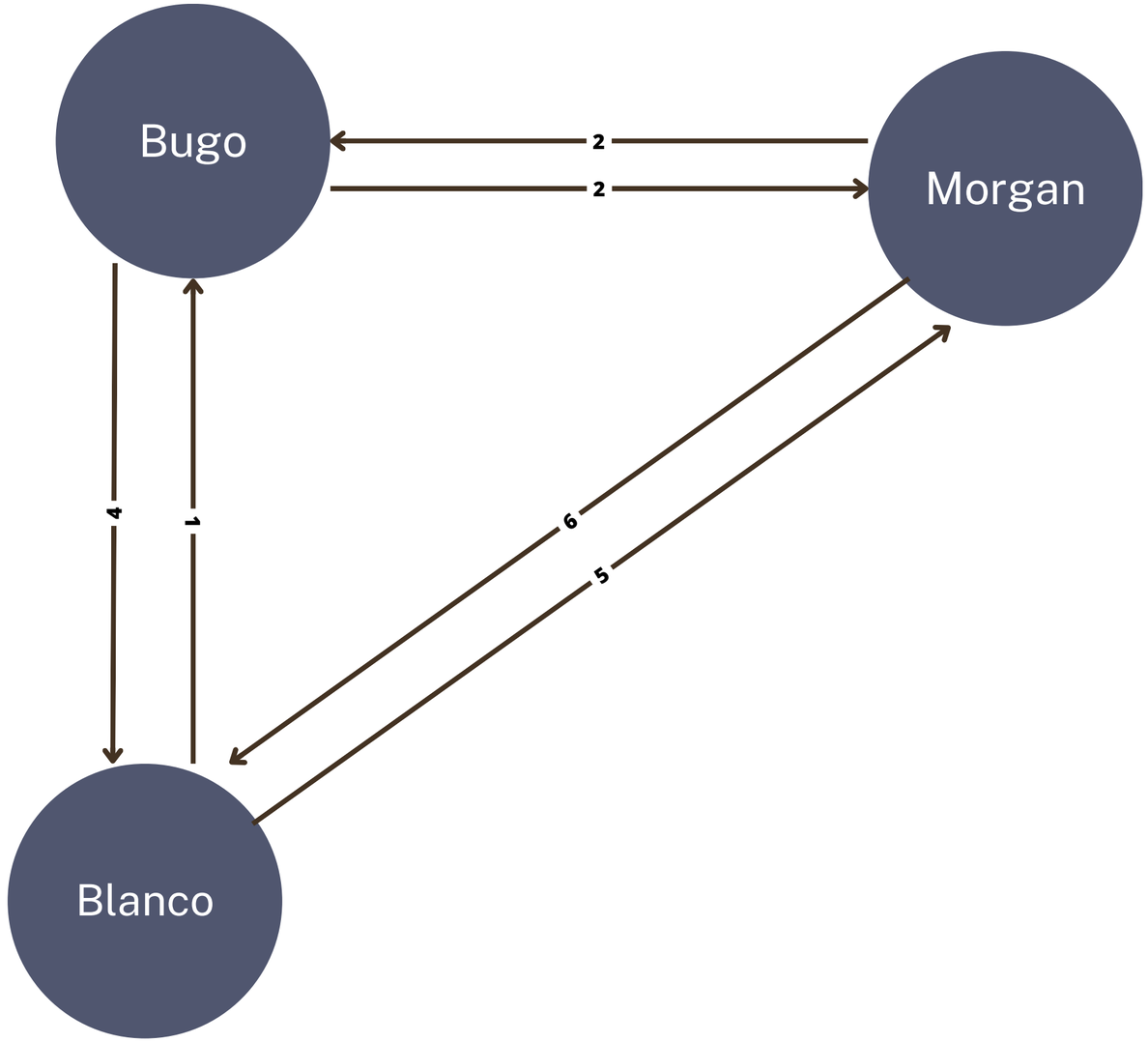
SIMILAR INTEREST COMMUNITIES (SIC)

Un gruppo di persone che rispondono allo stesso evento. Nel caso di Twitter, persone che spesso retwittano o rispondono agli stessi tweets. Questo implica che si condividono gli stessi interessi



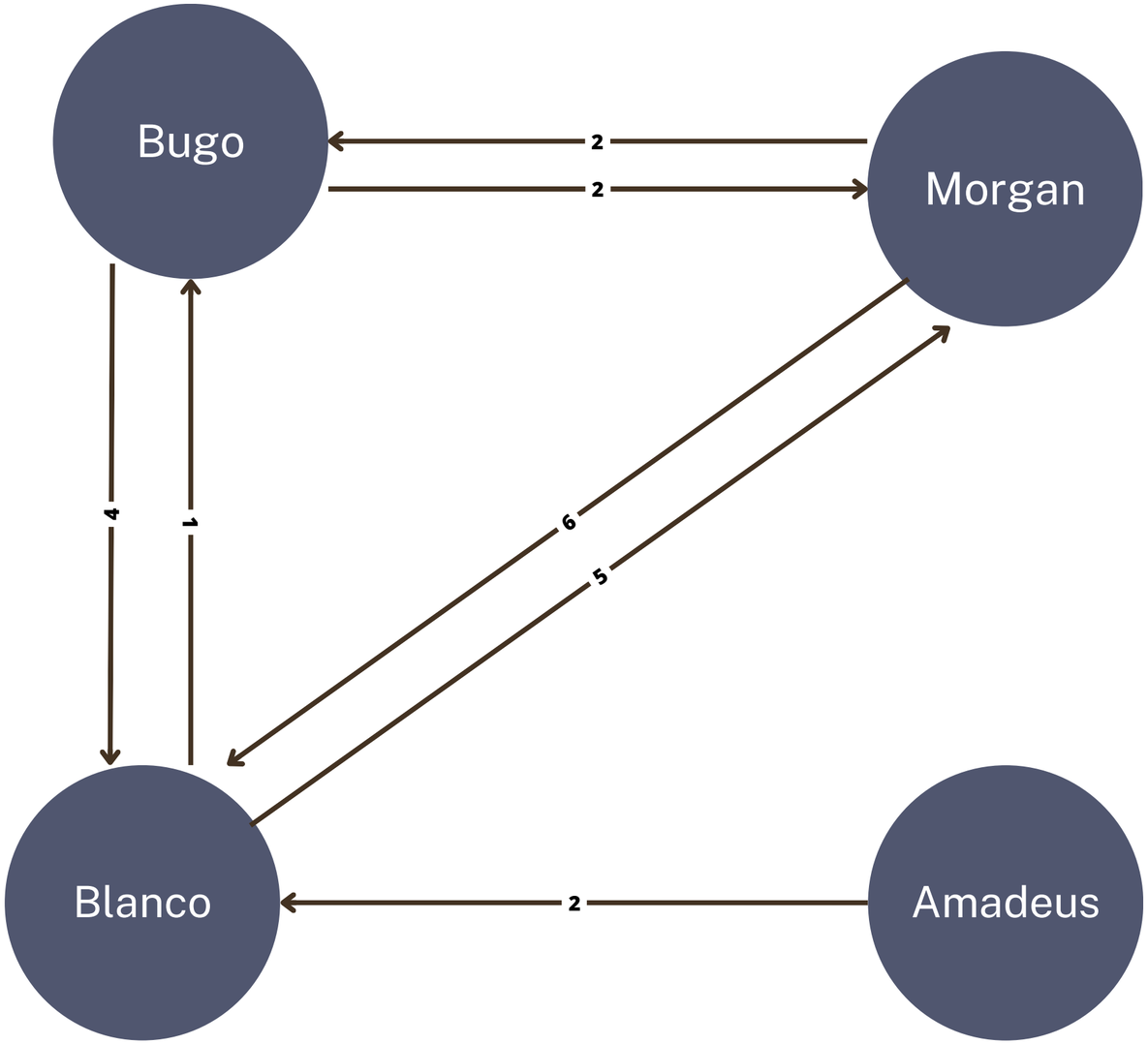
STRONG-INTERACTING COMMUNITIES (SC)

Un gruppo di persone che interagiscono tra loro.



STRONG-INTERACTING COMMUNITIES WITH THEIR “INNER CIRCLE” NEIGHBORS (SCIC)

Un gruppo di persone che interagiscono tra loro a cui si aggiungono membri che hanno conttatti con elementi del gruppo

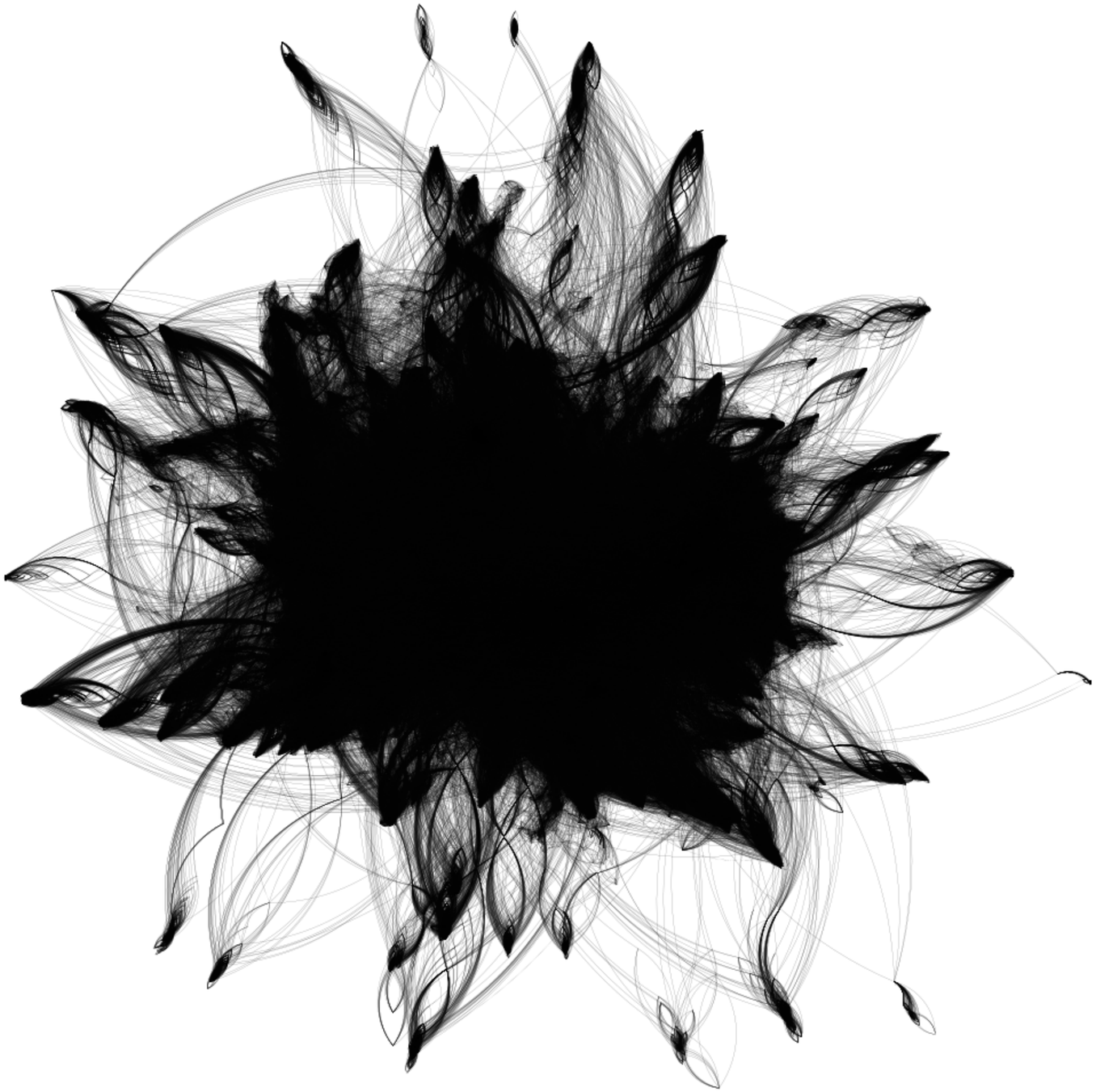


DATASET

Inteerazioni tra utenti su twitter. Ogni utente è identificato da un ID. Ogni nodo è un utente, ogni arco da A a B modella la citazione di B in un tweet di A

Il grafo contiene:

- 81306 nodi
- 1768149 archi

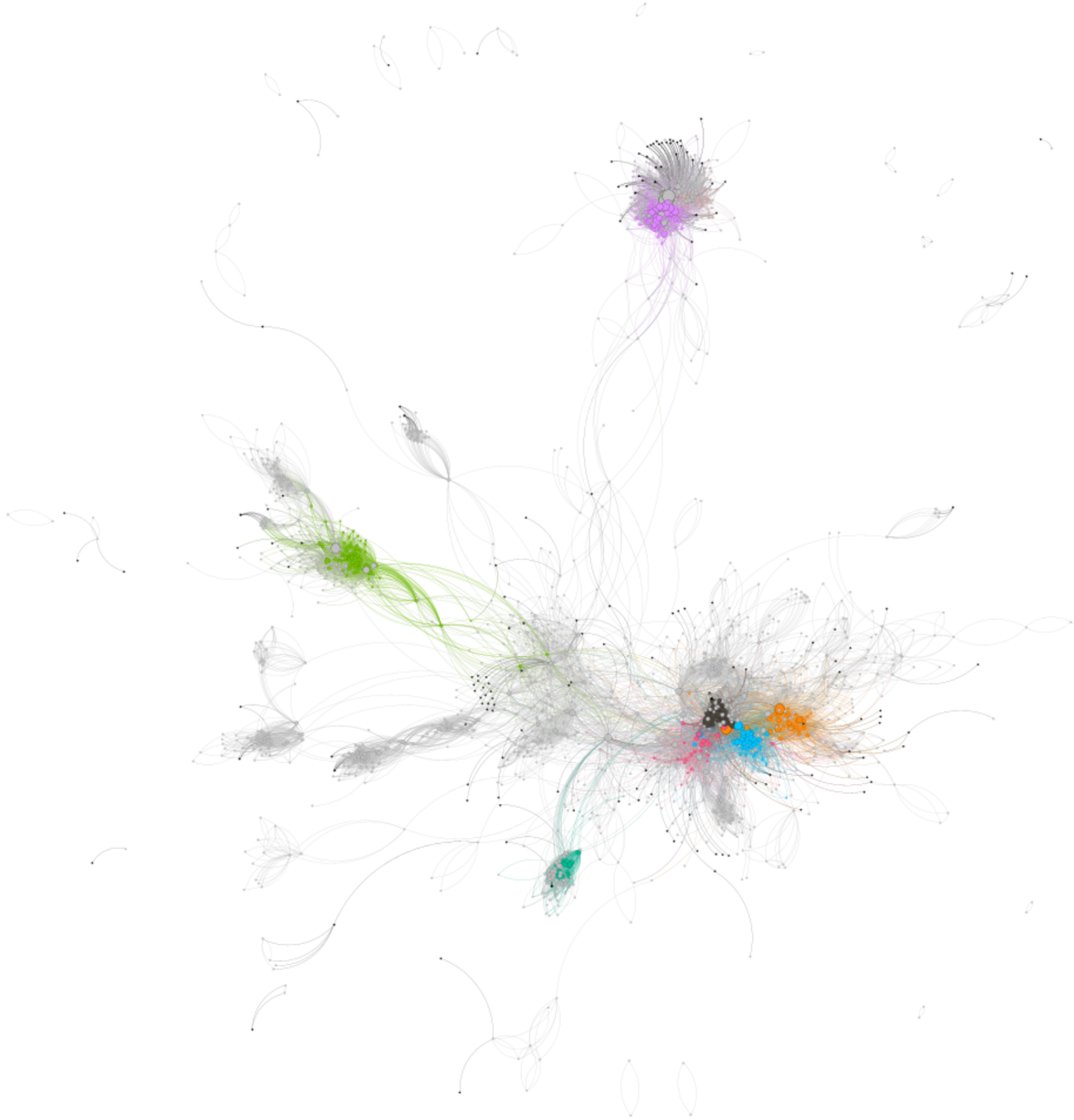


ALGORITMO SIC

Sono emerse 8 community che coprono 410 nodi su quasi 2000 nodi. I restanti nodi, con questa configurazione, non appartengano a nessuna community.

Input params:

- weight threshold = 5
- inDegree threshold = 3

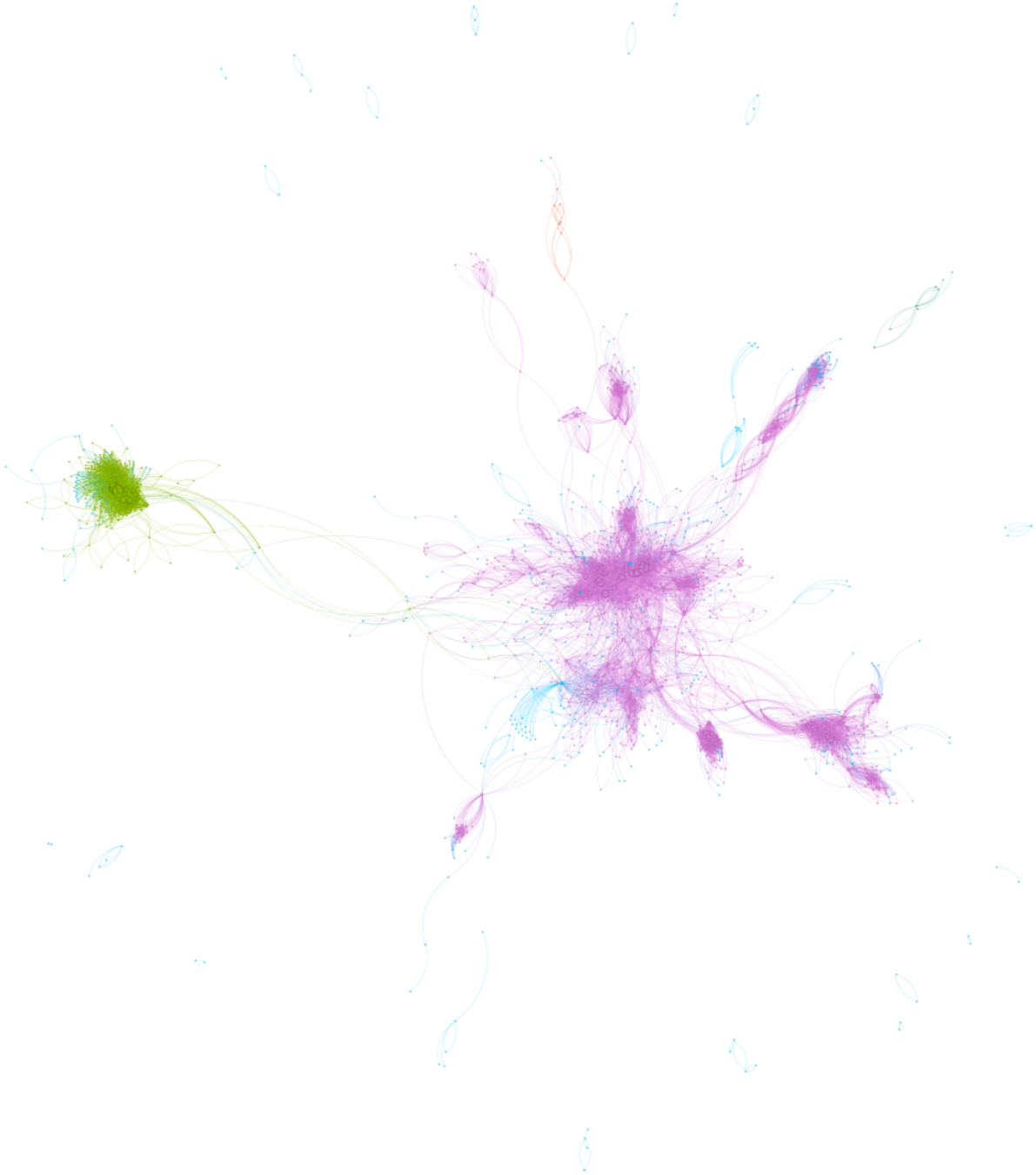


ALGORITMO SC

Sono emerse 5 community che coprono tutti i nodi. Le community risultano in parte sovrapposte. Questo è uno dei problemi di questi algoritmi.

Input params:

- weight threshold = 5
- inDegree threshold = 3
- min element per com = 5

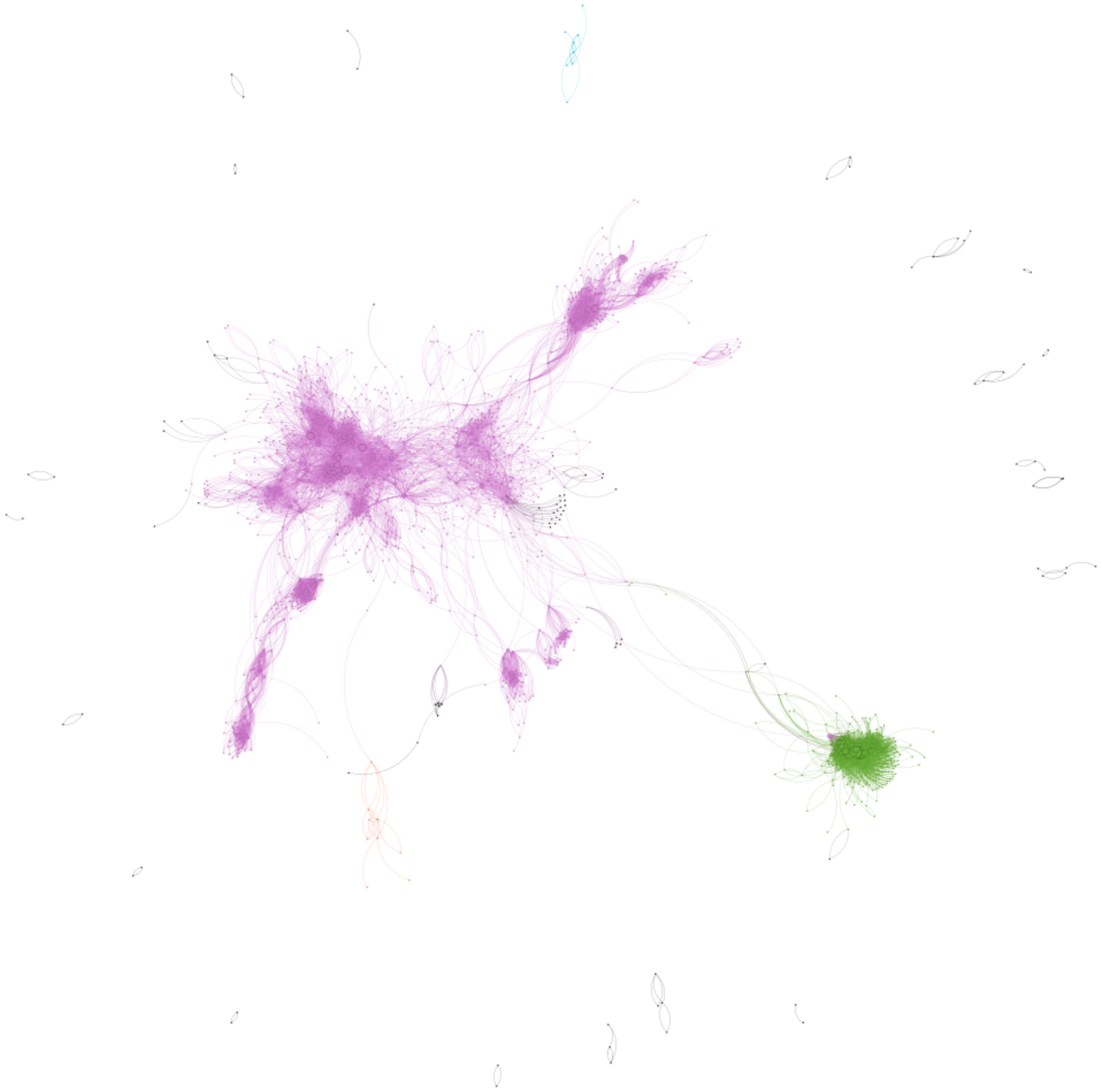


ALGORITMO SCIC

Sono emerse 4 community in parte sovrapponibili a quelle rivelate alle precedenti.

Input params:

- weight threshold = 5
- inDegree threshold = 3
- min element per com = 5



NOTE

CODICE E RISULTATI

Repository Github

Grazie per l'attenzione