Almawave

Structured data mining on Twitter datasets

MENTOR
ING. CRISTINA GIANNONE

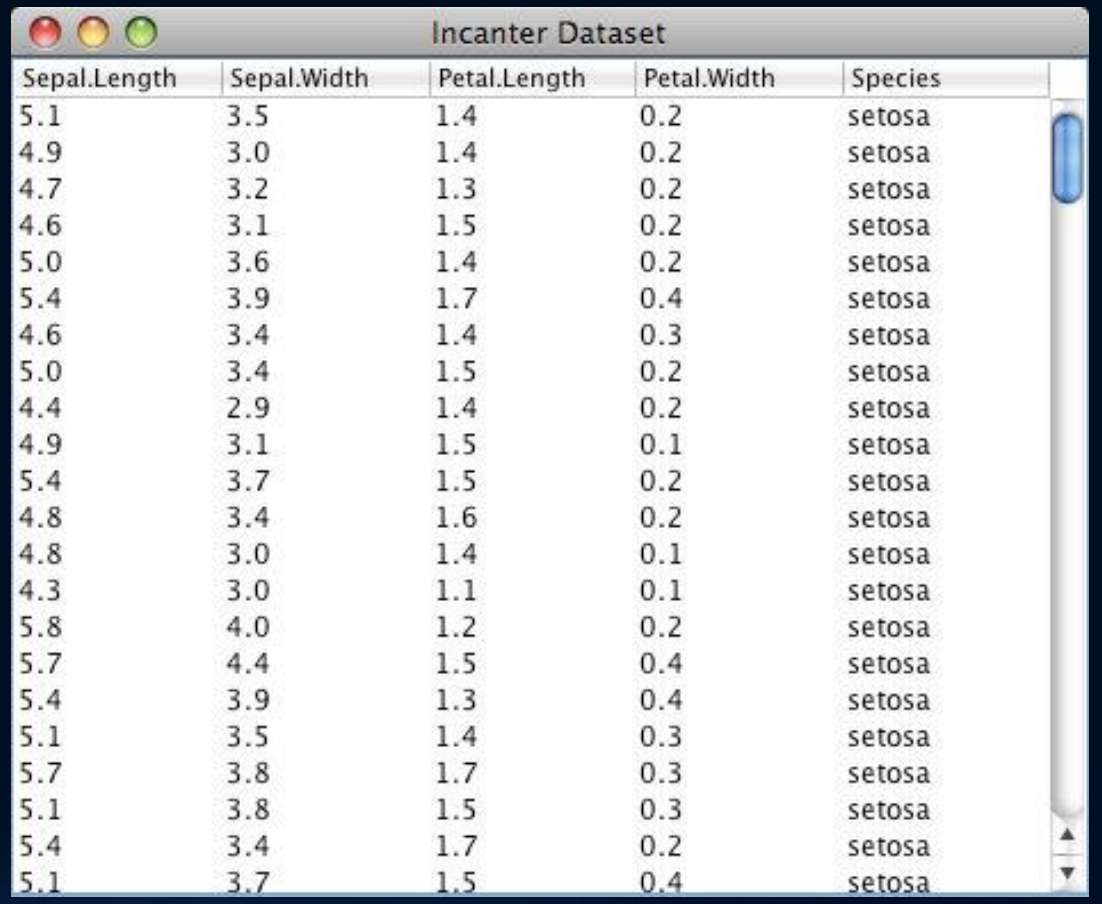
INTERNS
A.GALLO, S.TILIA, W.RUKUN

Introduction

Glossary

Glossary

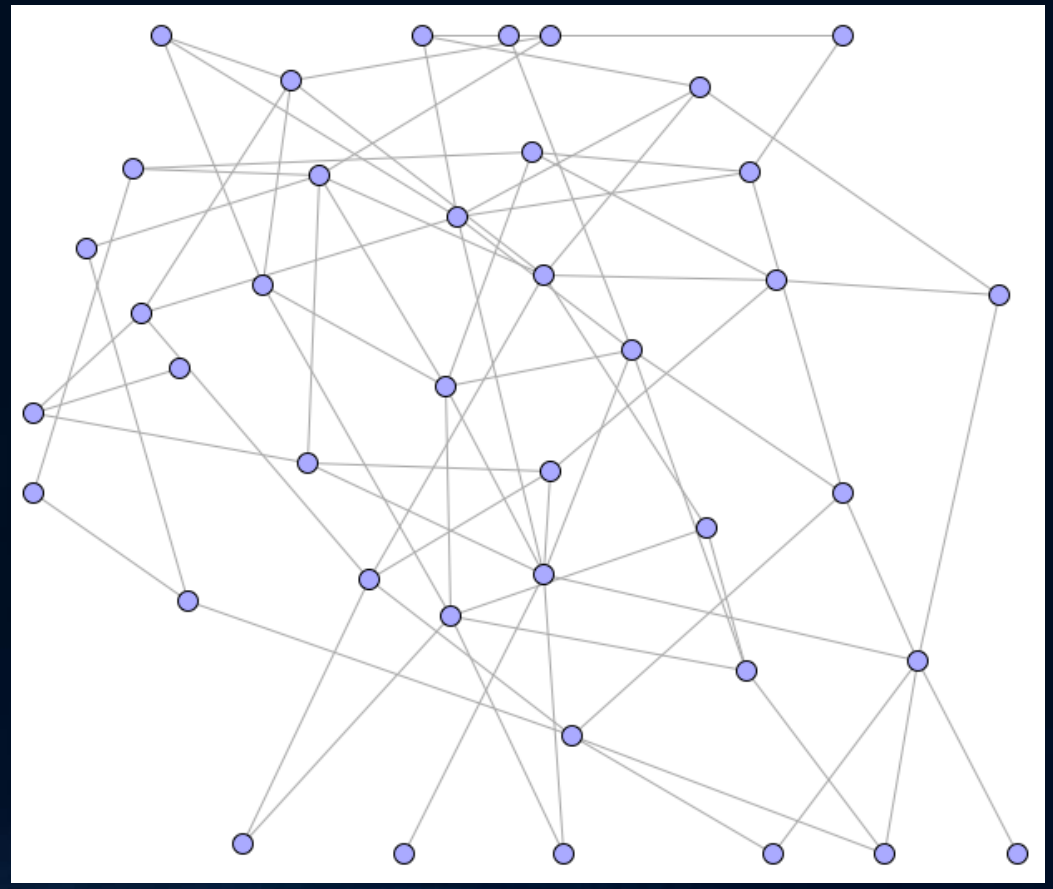
- **Dataset:** collection X of n elements with i attributes



Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa
5.4	3.4	1.7	0.2	setosa
5.1	3.7	1.5	0.4	setosa

Glossary

- **Graph:** structure amounting to a set of objects related with nodes and edges

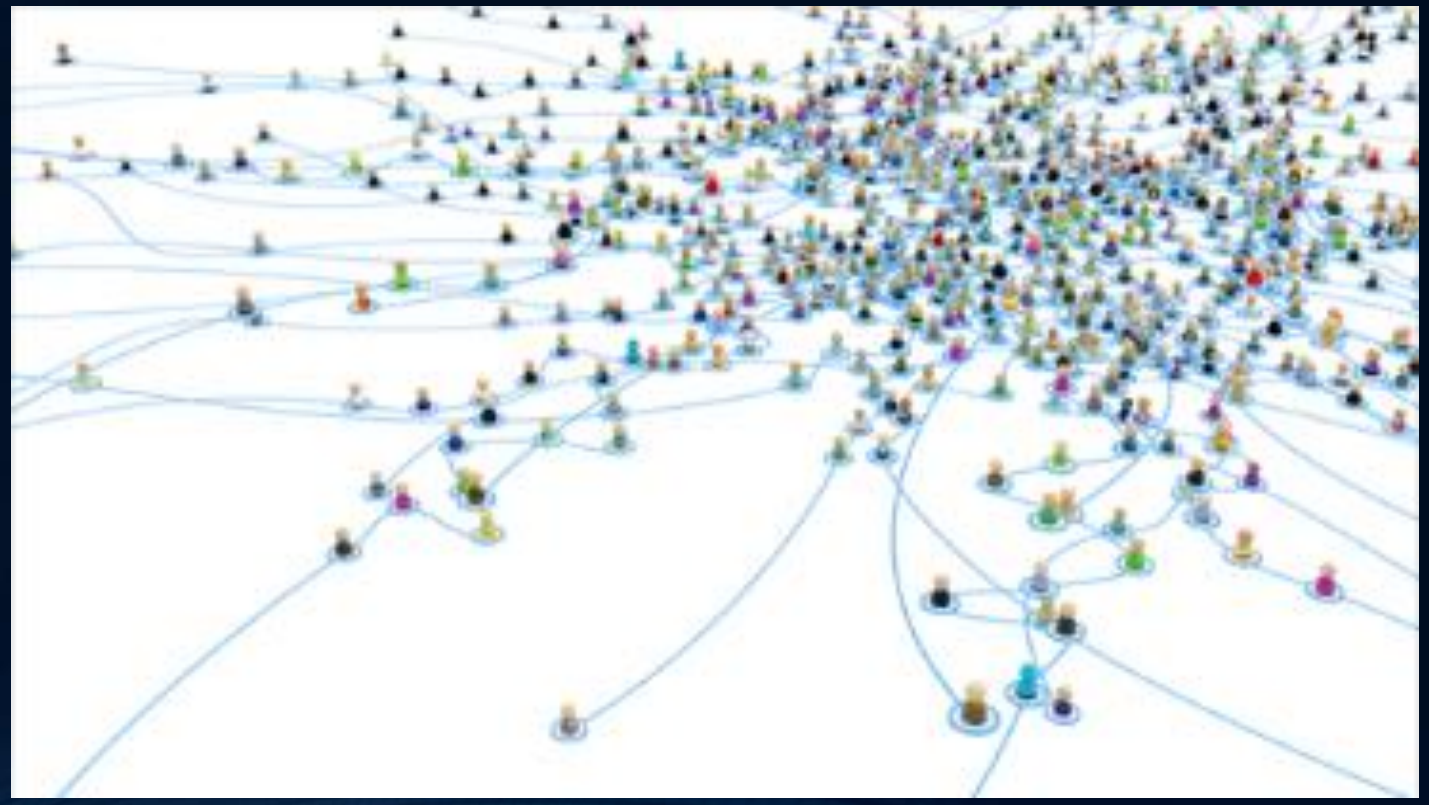


Glossary

- **Relation:** structure amounting to a set of objects related with nodes and edges

Glossary

- **Social network:** heterogeneous and multirelational dataset represented by a graph, with nodes (or vertex) corresponding to objects and edges corresponding to links representing relationships or interactions between objects



Glossary

- **Twitter:** social network where people, called users, can interact with "tweets", a message of 140 characters. They can "retweet" them or tag other users

Tools

- **Programming language:** Python
- **Architecture Tool + Platform:** Hadoop + Spark
  - **Hadoop**: Application programming interface centered on a data structure called the resilient distributed dataset (RDD), a read-only multiset of data items distributed over a cluster of machines

    Spark: Open-source software framework used for distributed storage and processing of very large data sets. It was developed in response to limitations in the MapReduce cluster computing paradigm
- **Notebook + Platform mod:** Ipython + Pyspark
  - Ipython: command shell for interactive computing in multiple programming languages, developed for the Python. Pyspark: exposes the Spark programming model to Python
- **DataFrame Tool Package:** Graphframe
  - GraphFrames is a package for Apache Spark which provides DataFrame-based Graphs and represent graphs with vertices (e.g., users) and edges (e.g., relationships between users)

The aim of the project

What's the question?

- How can we determine communities on a social network like Twitter?
- What can we do then?

Twitter structure

Relations between tweeters

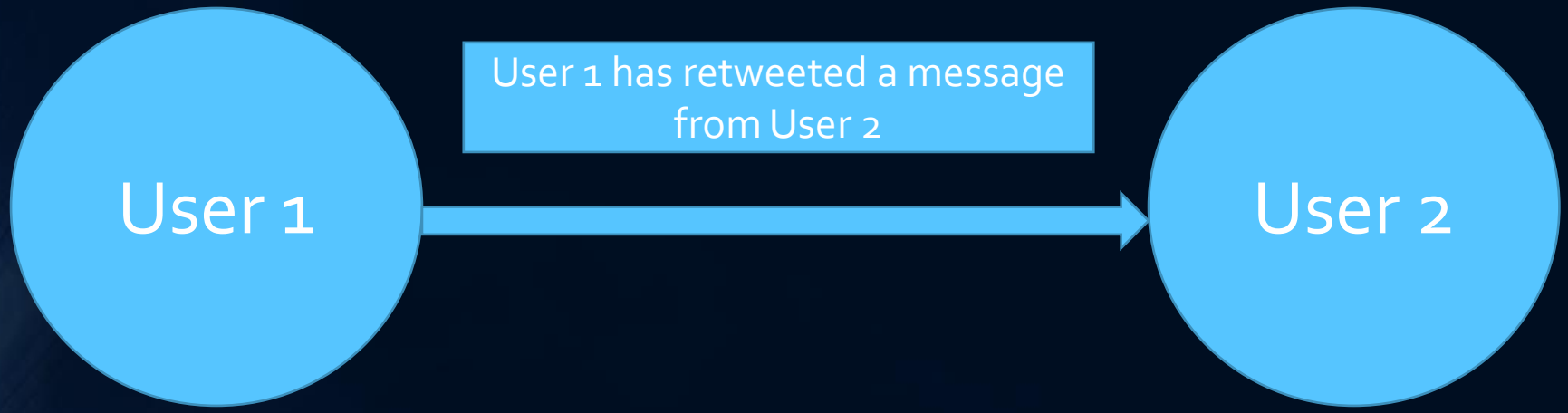
How to choose the nodes and the edges of the graph?

We thought that the most important interactions between the users are three

Relations between tweeters: retweet

How to build a graph with retweets?

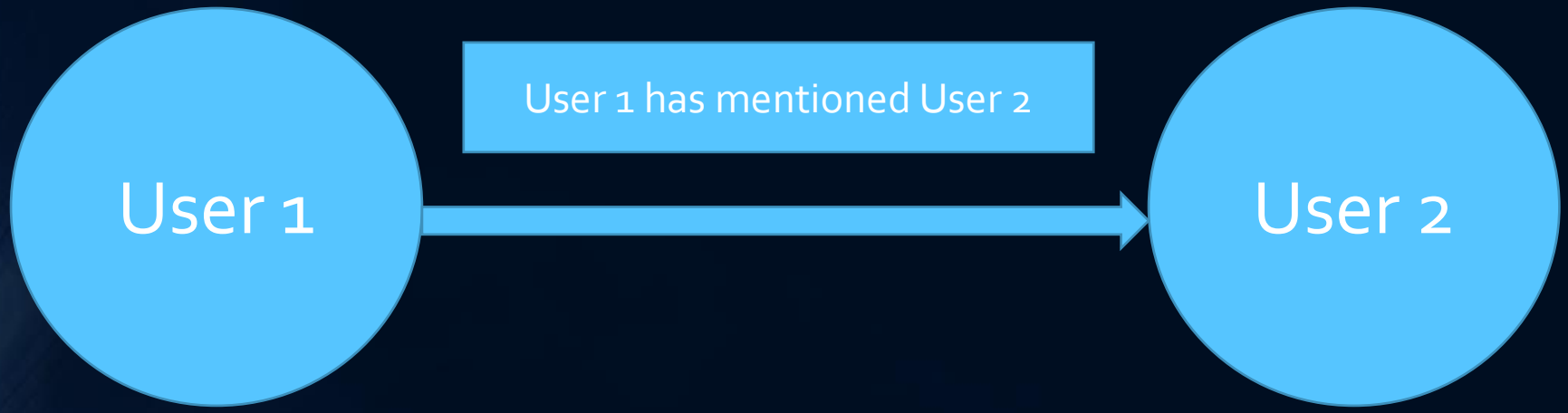
Put an edge from an user (called source node) to another user (called destination node) when the first user retweets a tweet from the second user



Relations between tweeters: mentions

How to build a graph with mentions?

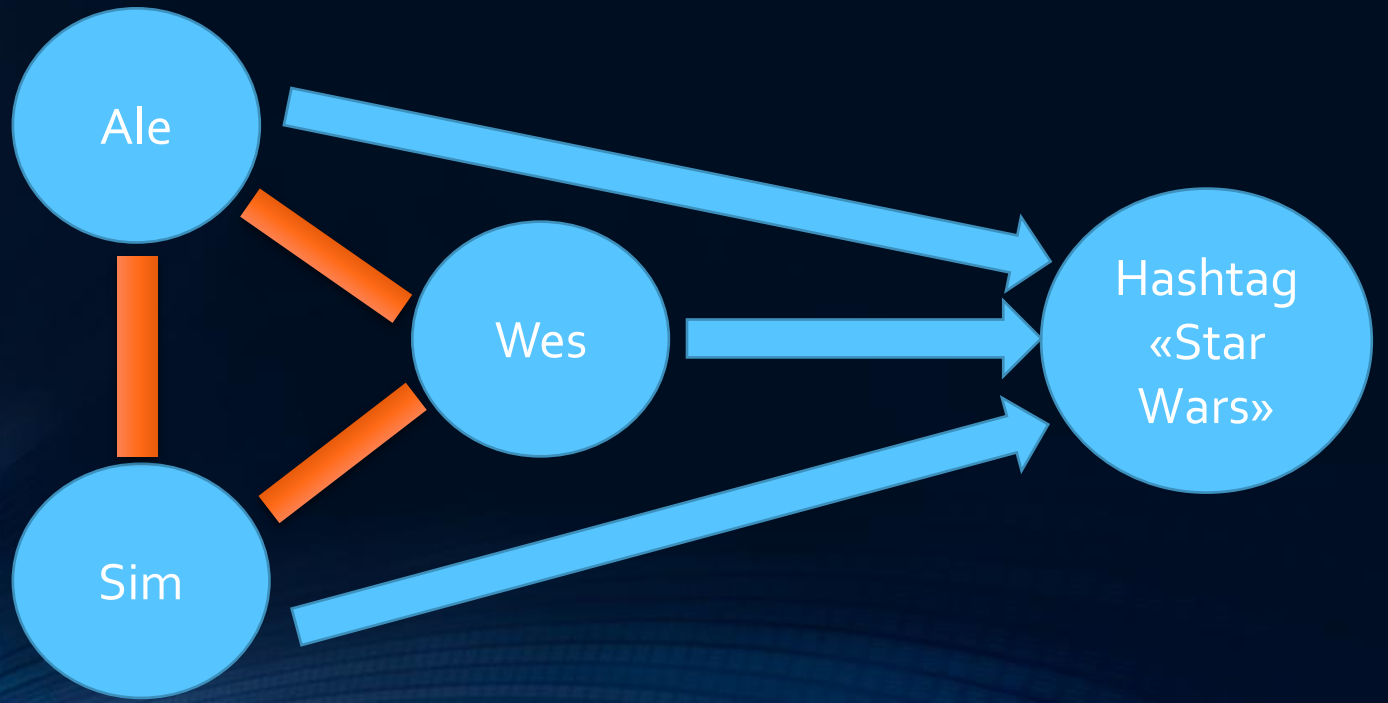
Put an edge from an user (called source node) to another user (called destination node) when the first user mentions the second user



Relations between tweeters: hashtags

How to build a graph with hashtag?

1. Put an edge from an user (called source node) to a hashtag (called destination node) when the user mentions the second user
2. Create a new graph putting an edge between all the users who shared (were connected to) the hashtag

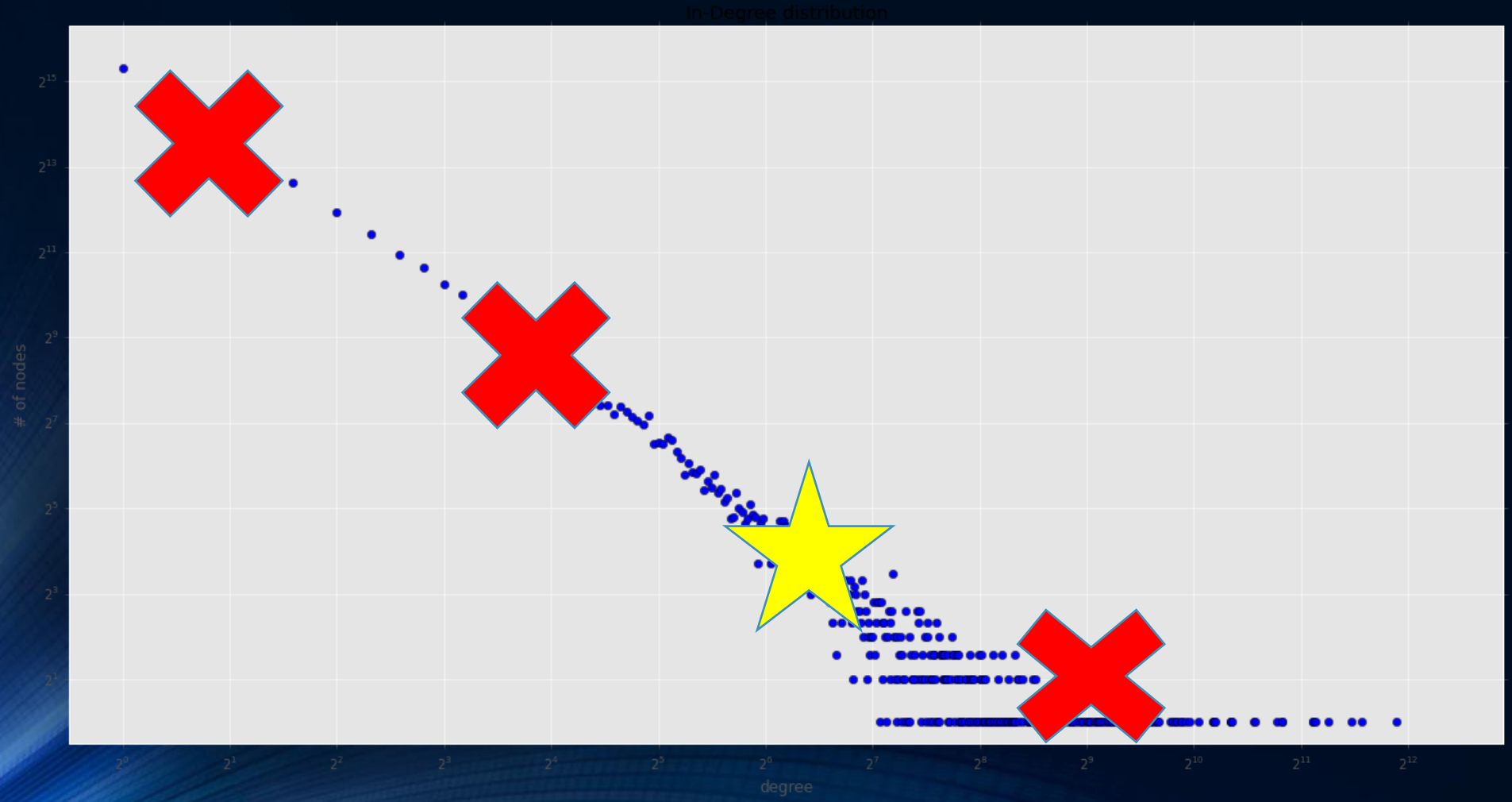


Our approach

- We decided to choose the 3rd graph to make a model built on the common interests of the users.
- For instance: Alessandro, Simone and Wesam tweet the hashtag "Star Wars". They are a graph with edges directed to the node/hashtag.
- We build so a *new graph* with all the nodes that shared the same hashtag so Alessandro, Simone and Wesam will be connected by undirected edges

Filter the graph

We made the 3nd choise but, before that, we needed to filter the hashtags that are too much or too few popular for the sake of real communities.



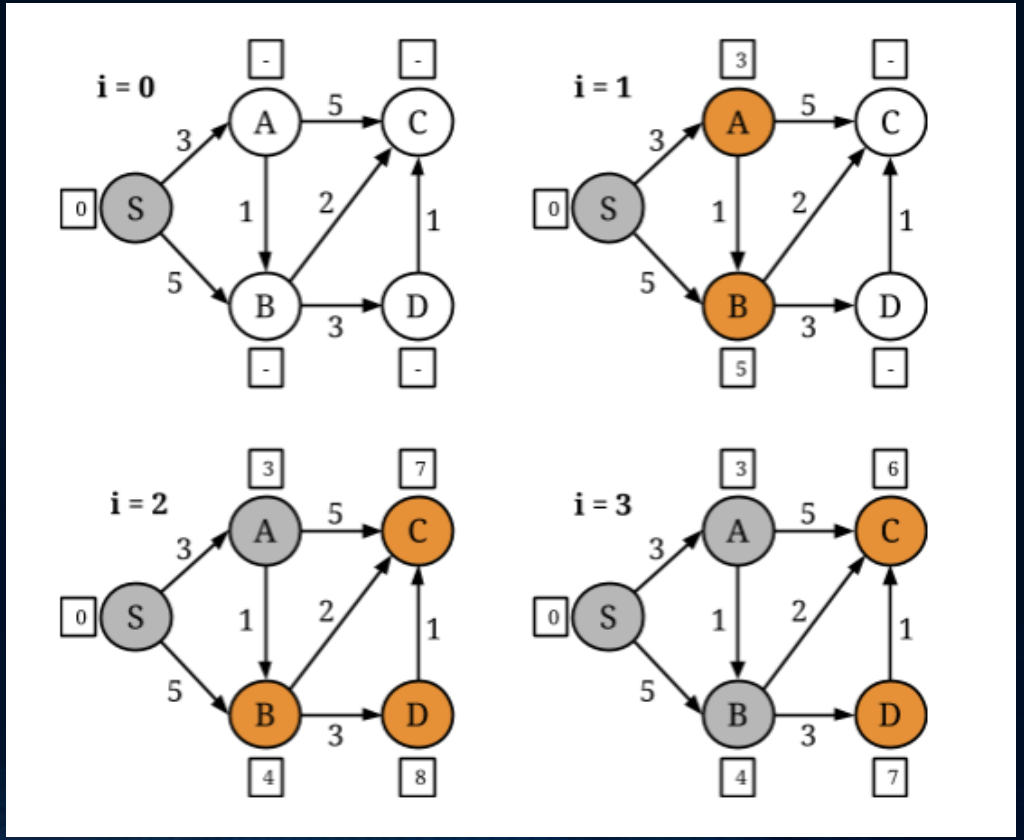
Algorithms

Page Rank

The Page Rank is an algorithm used by Google Search that works by counting the number and quality of links to a user to determine an estimate of how important it is. The underlying assumption is that more important users are likely to be connected to more users.

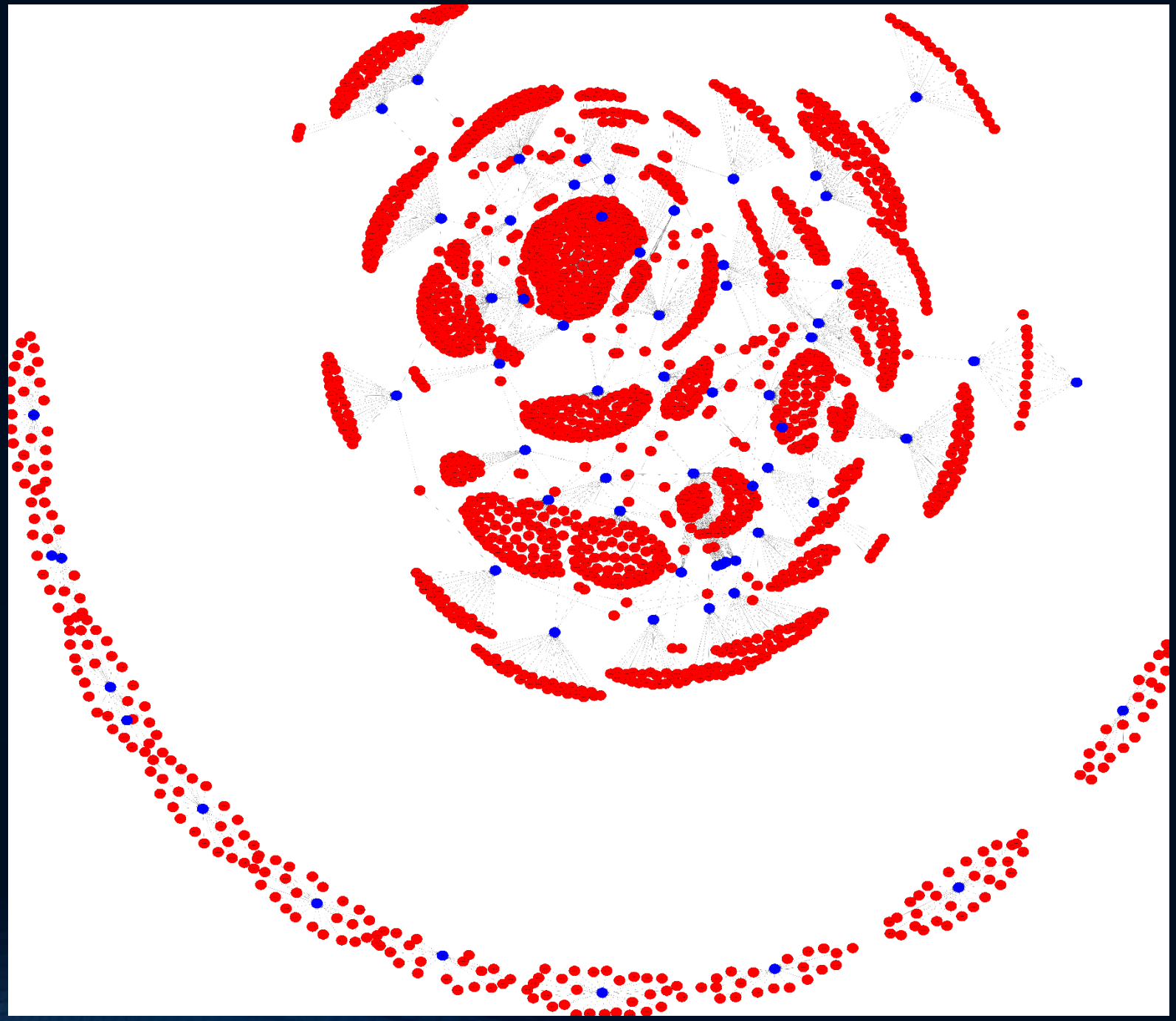
Label Propagation

Label Propagation is an algorithm used to find the communities on the networks. Each user in each iteration will get the most common label in its neighbourhood.



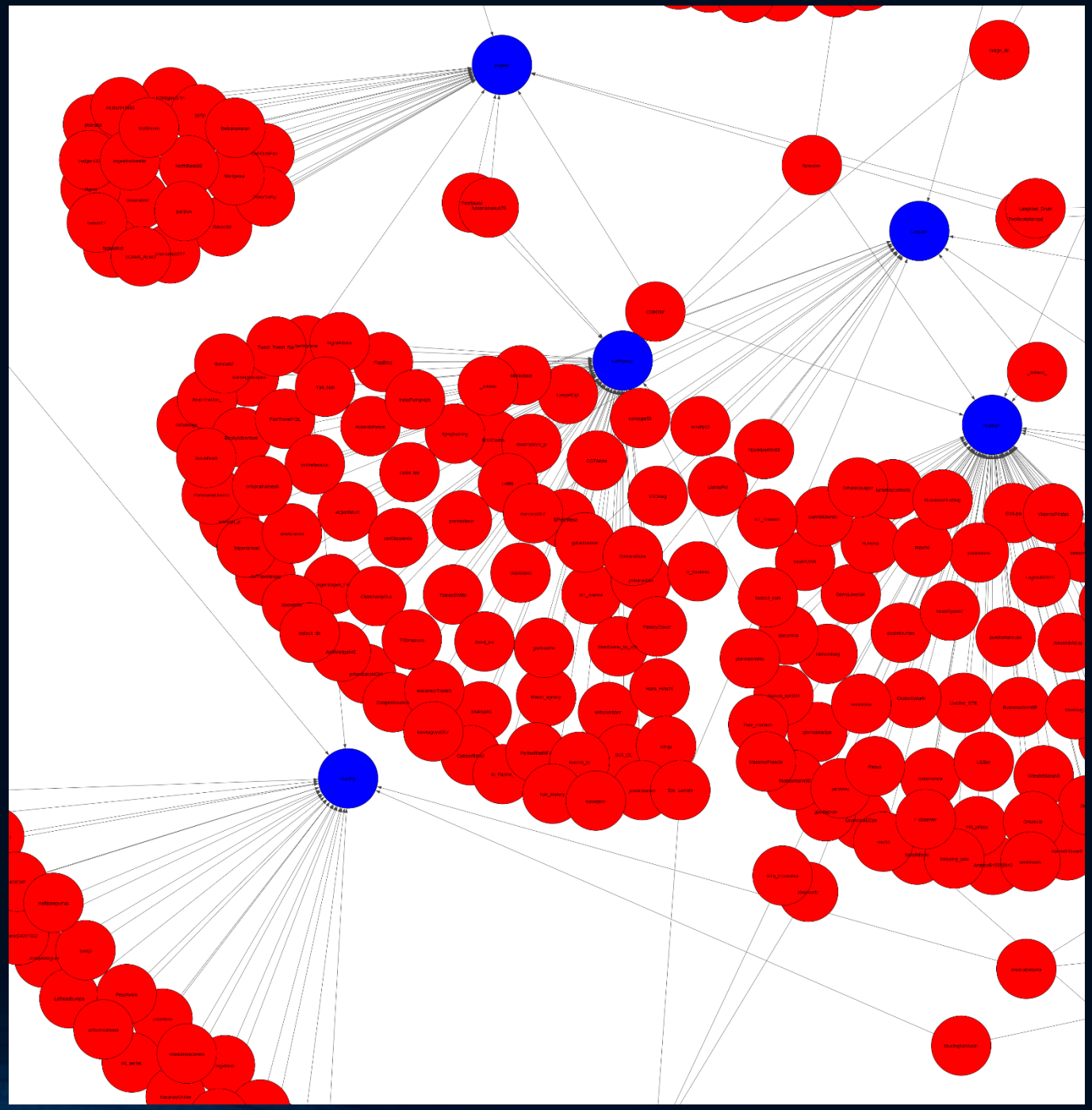
Results

Hashtag Graph

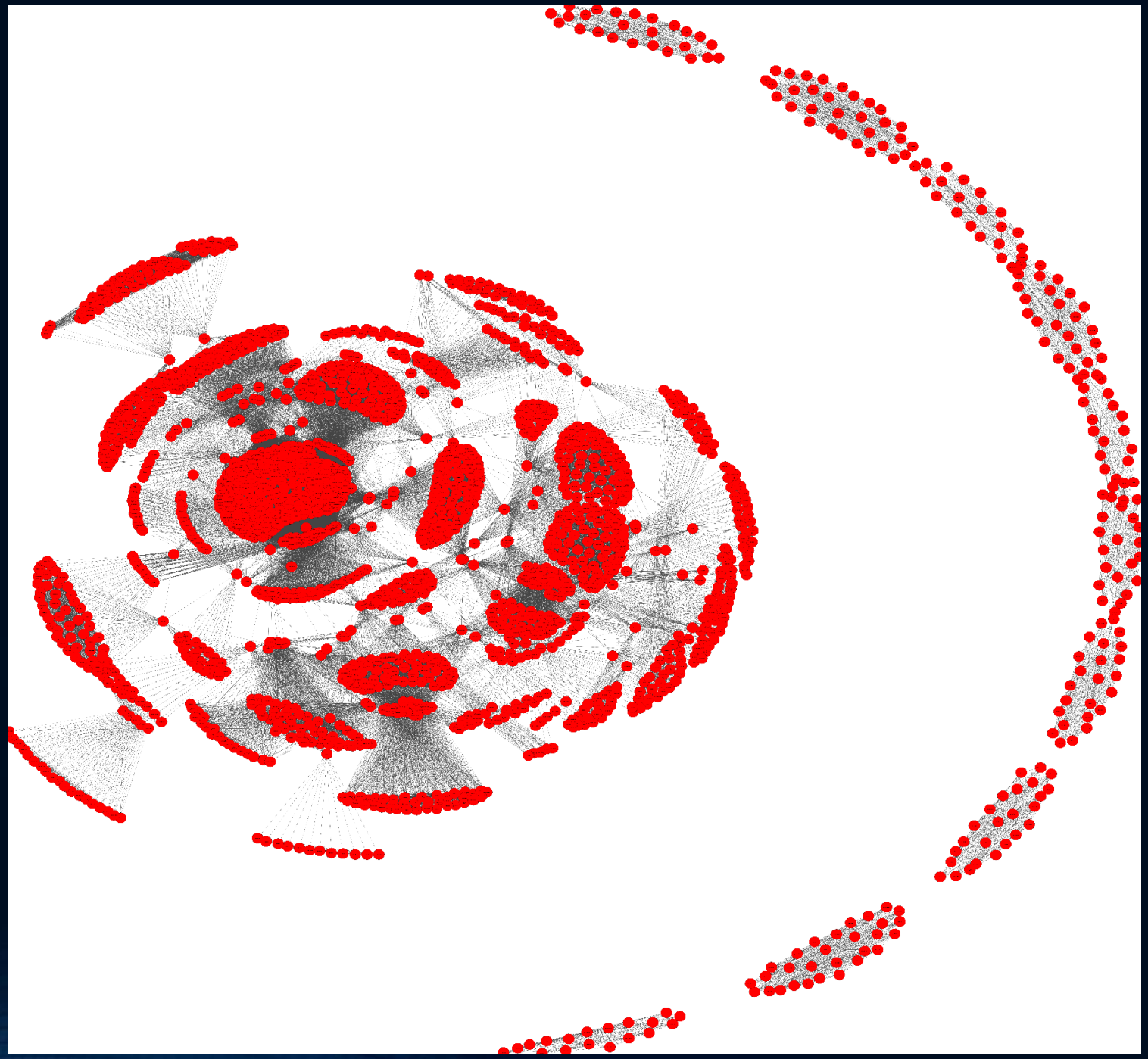


Hashtag Graph

(zoomed)



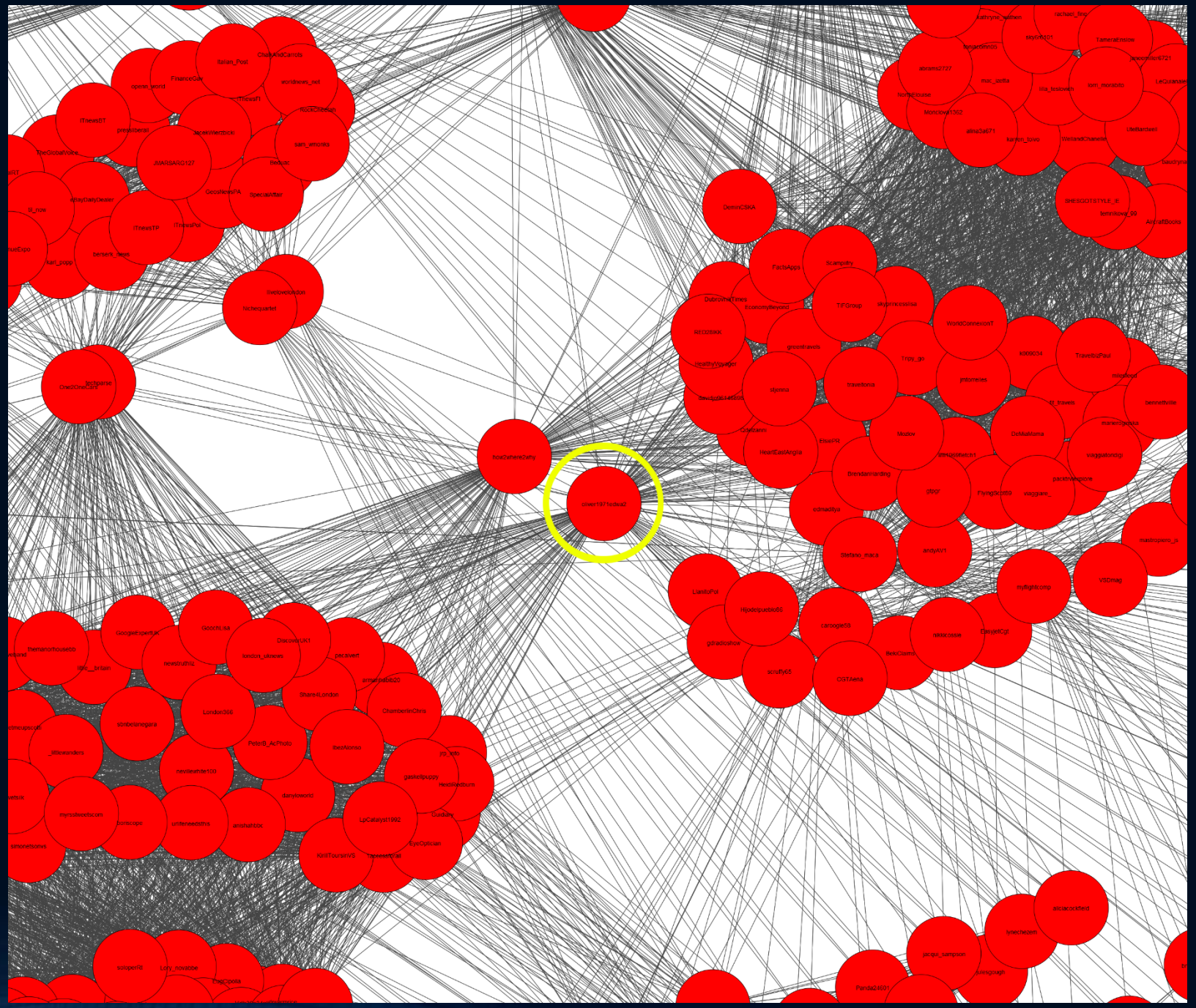
Users Graph



Users Graph

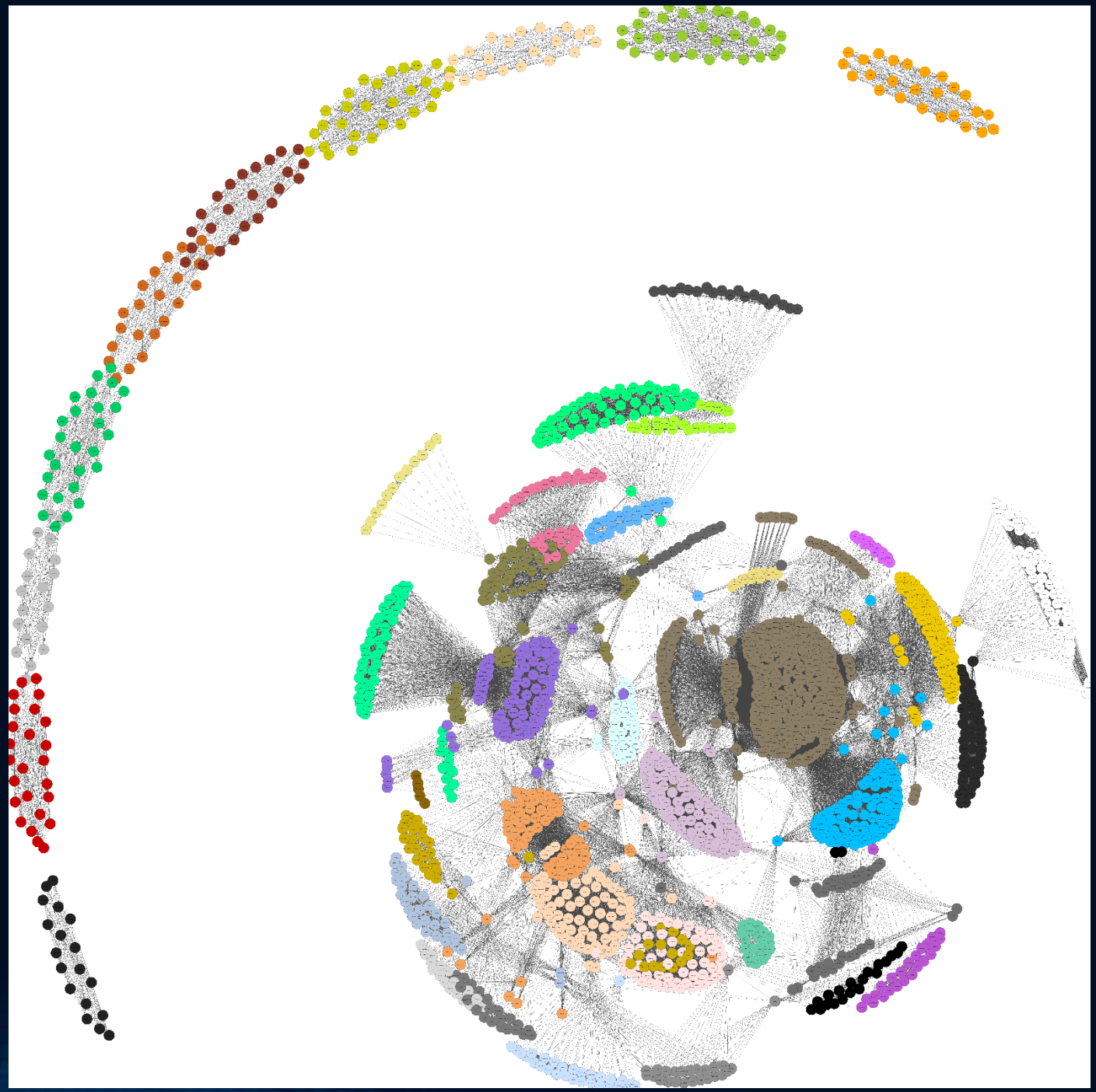
(zoomed)

- The yellow circle in the center is the most central one (“oliver1971edwa2”), according to the PageRank algorithm



Label Propagation

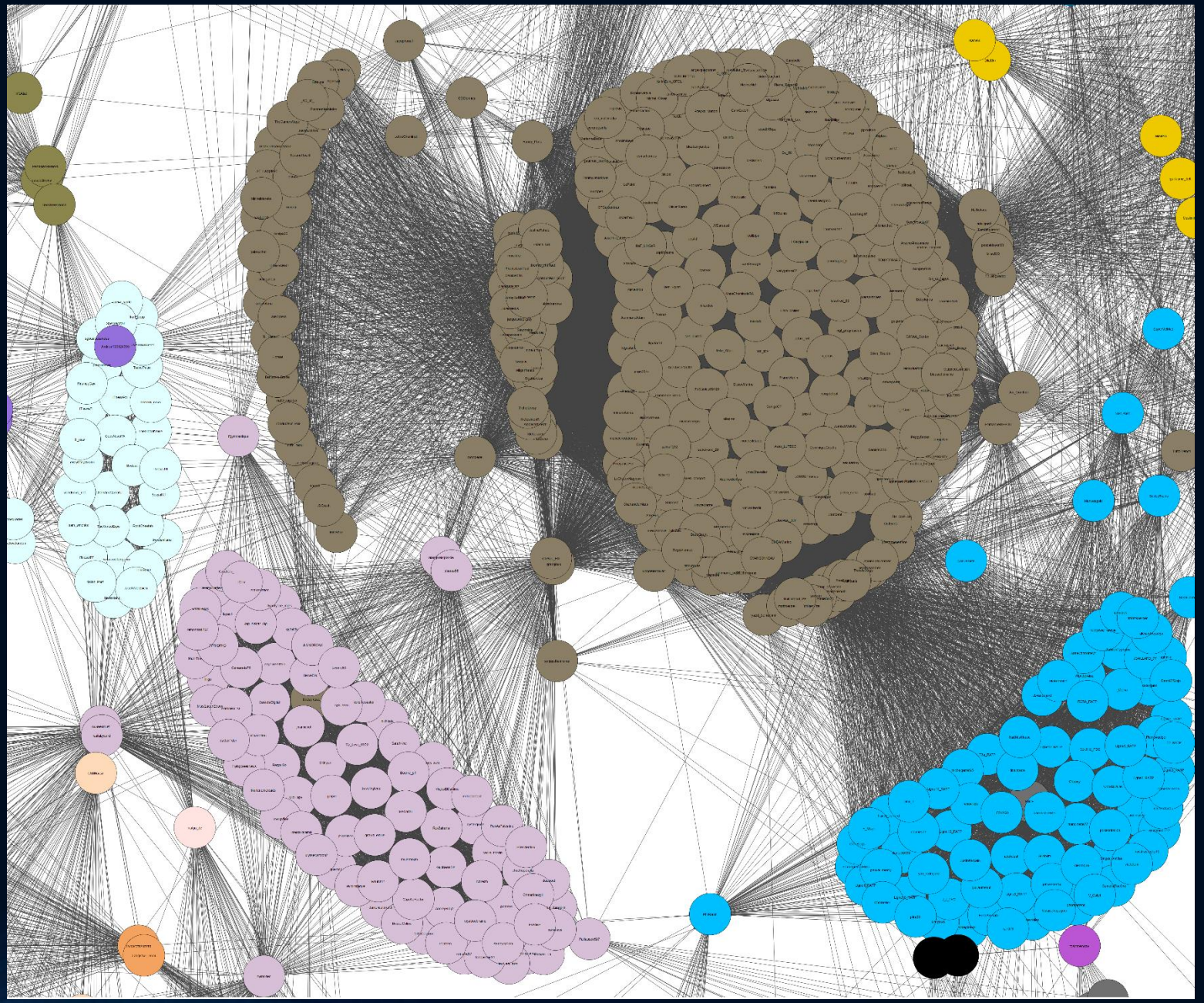
- The result of label propagation shows different colors for each different community.



Label Propagation

(zoomed)

- For instance, the central blue node (“PhilRouin”) in the down-right side of the image belongs to the blue community



Page Rank

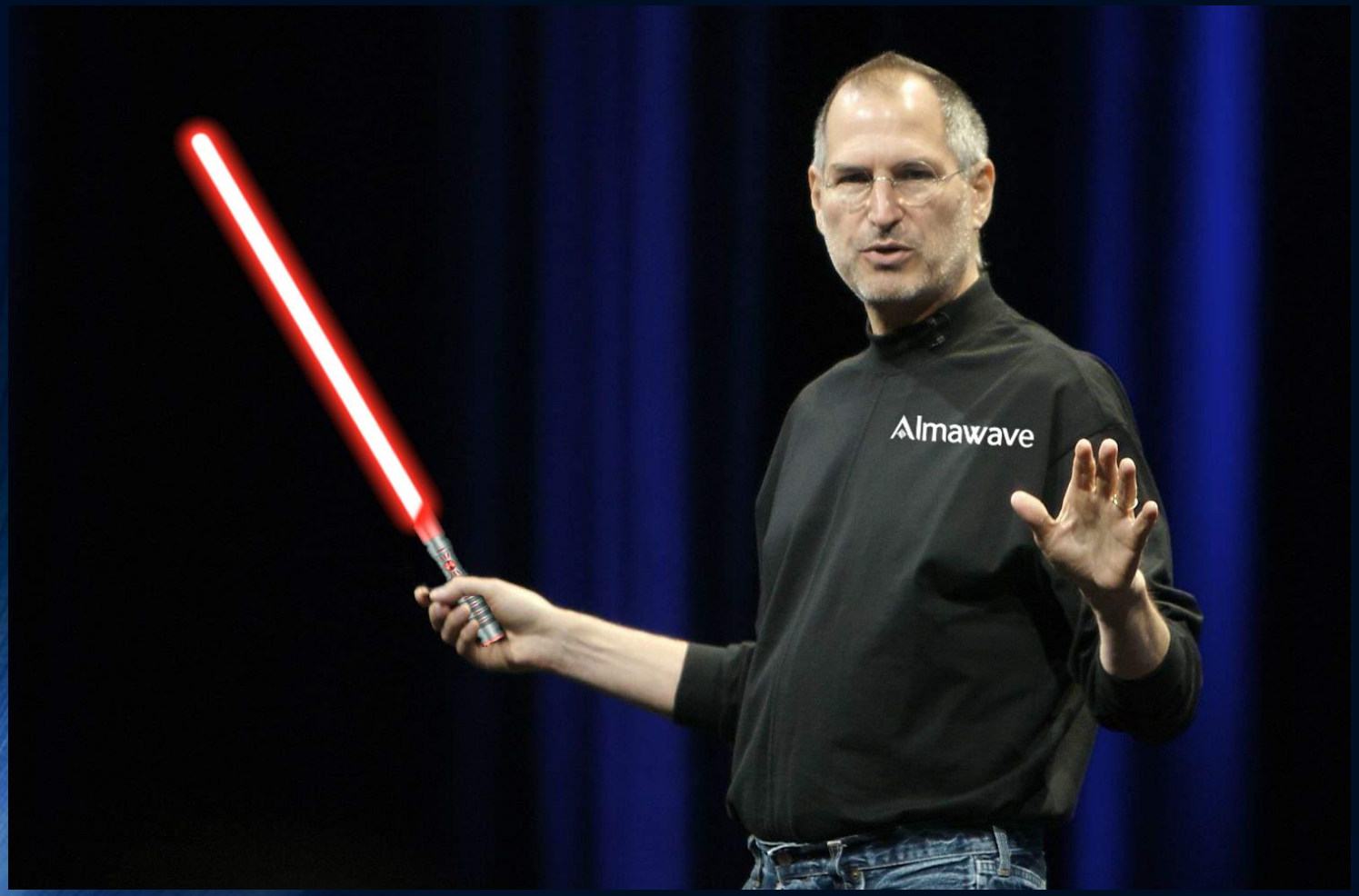
id	pagerank
oliver1971edwa2	2.568101267037709
flyinsider	2.3928569389282695
EmmanuelTEBOUL	2.0839782220753023
CDBDSF	2.0831474867326625
flight_base	1.9643466374485672
Stefano_maca	1.904216442218753
Mondeferroviair	1.8254444507778258
MonWagon	1.8217319468542867
grangibus	1.8112698164104517
CtrISec_FR	1.8112698164104517
lowcostinfo	1.7876964509641162
FeliciLory	1.7669453611139154
Jon_Gonthier	1.7234675396113632
TM_CAPLAIN	1.7225811089617056
OuthierC	1.7225811089617054
GeosNewsRoma	1.7154292261390751
temnikova_99	1.7090956194365152
AircraftBooks	1.7090956194365152
SHESGOTSTYLE_IE	1.7090956194365152
sindicatostavla	1.7004444610009408

only showing top 20 rows

These users are the most central ones according to the PageRank algorithm.

- It means that they are users with several interests and they tweet on hashtags that are relevant but not so famous.
- We could sell a product through them instead of using popular influencers.

Page Rank algorithm



For instance, we'd like to sell lightsaber.

- We consider many of these potential influencers that could be interested and pay them to advertise our amazing laser sword.
- They start to tweet about it and the product could reach so several communities.
- ***BIG MONEY TIME***.

Thank you for your attention