Community detection and labeling in an Instagram Network

Ahmed Furkan Ozkalay
Eda Bayram
Orçun Gümüs
Osman Berk Satir

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

A Network Tour of Data Science, Fall 2017
Prof. Pierre Vandergheynst
Prof. Pascal Frossard

Motivation

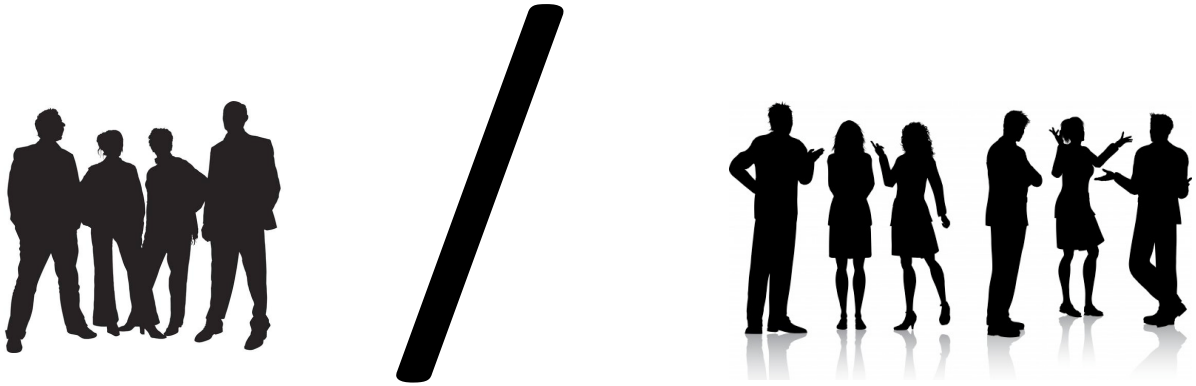
- Analysis of social networks:
  - Detecting community structure (Girvan and Newman, 2002)
  - Finding common property representing the community → Labeling
    - ”Like” information
    - Followed pages
    - Followers
    - Visited places
    - Bio (Age, Sex, Job)

Objective

- Community detection by Spectral Clustering
- Exploit multiple endpoint information of users
  - Acquisition of the network → “Like” relations, location
  - Constructing the graph → “Like” relations
  - Labeling → Visited places
- No correct way of separating clusters
  - Visualize inter & intra connection
  - Finding common location
  - Comparison with other methods

Data Acquisition and Cleaning

We have decided to use social media data from Lausanne area to make our findings meaningful at the same time interesting.



Which data source

Graph API 2.0

+ Easy to use, consistent and robust

+ Provide places with their geolocation

+ Provide id mapping for Instagram API

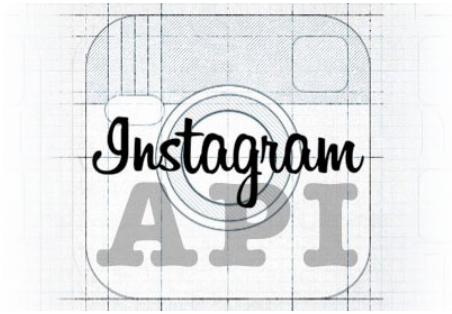
- Provide no relations in between users

Google Places API

+ Easy to use, consistent and robust

+ Provide places with their geolocation

- Provide no relations in between users



+ Provide in between user interaction via likes

+ Provide visited places of the users

- Inconsistent, some endpoints giving different results for the same parameters

Network

What do we need to construct a network?

- nodes - Information
- edges – Relations

Instagram Users

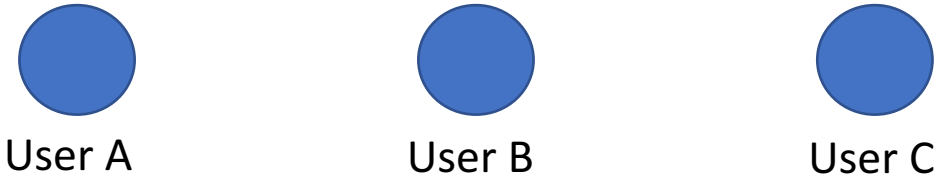
Visited

Liking posts

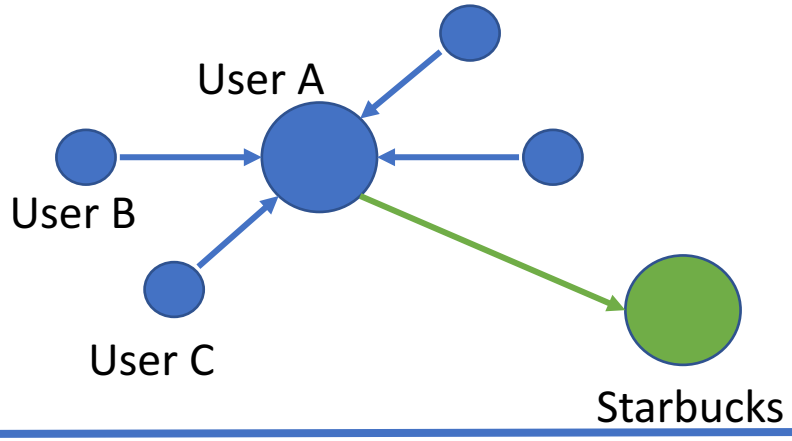
Visiting Same Place

Instagram

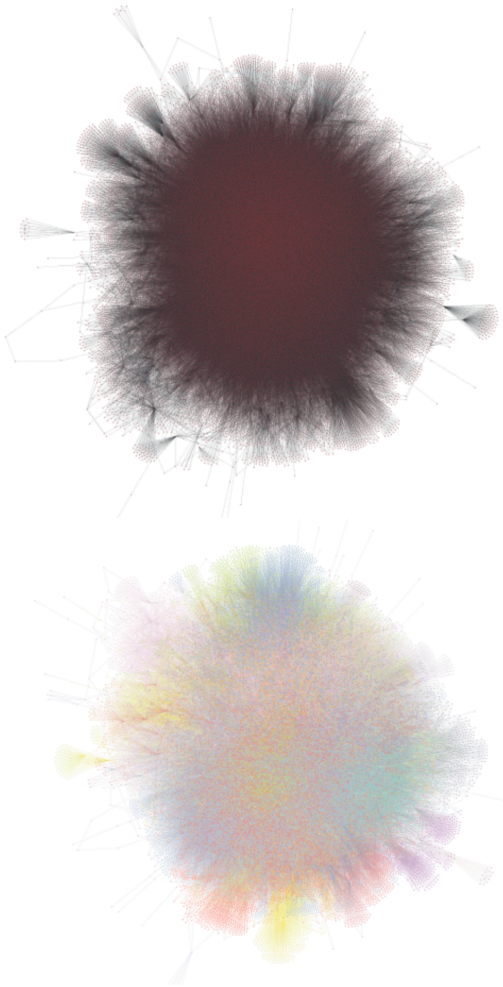
- Get users who share posts on any Lausanne place



- Enlarge them by their likes on his/her posts



Collection Order



We used Graph-tool which have a C++ backend to process large graph

2700 places

41000 users

283301 LIKE edge

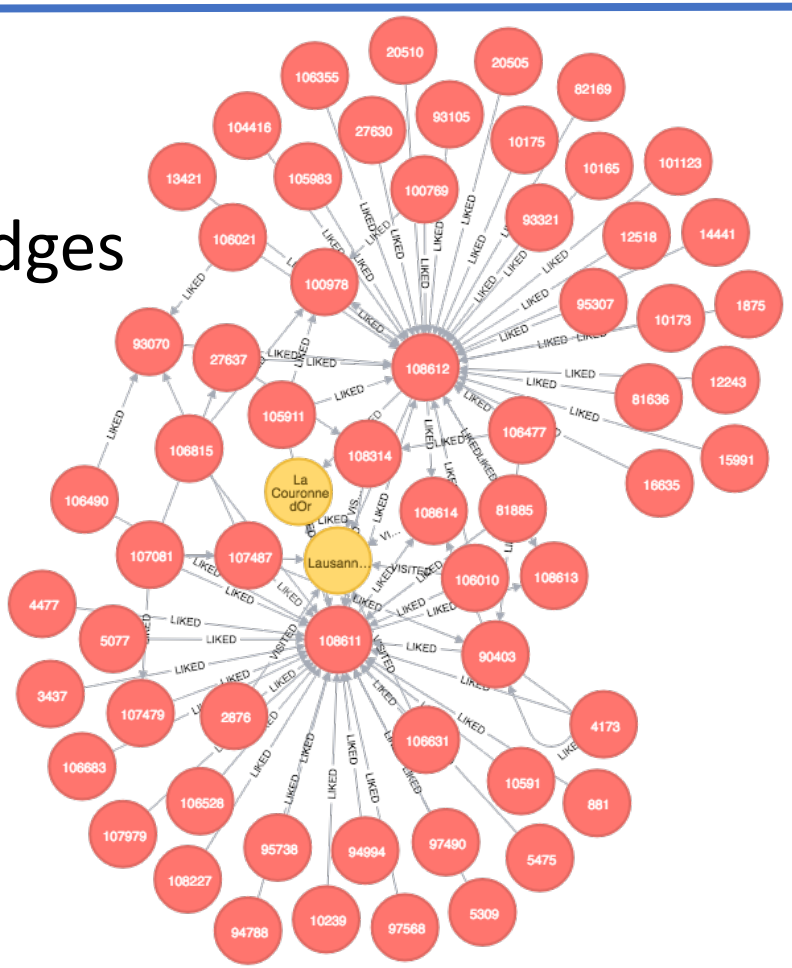
73000 VISIT edge

Purple nodes are users
Green nodes are places

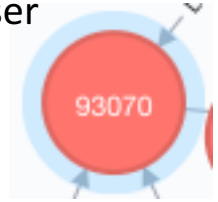
Data Format on Neo4j

visited edges

like edges



Instagram User



Facebook Place

We are decided to use Neo4j for our collected data from these APIs as well as our generated results to compare them easily.

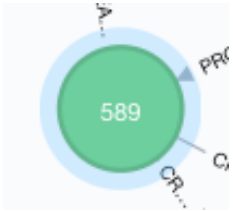
Saving Community Results

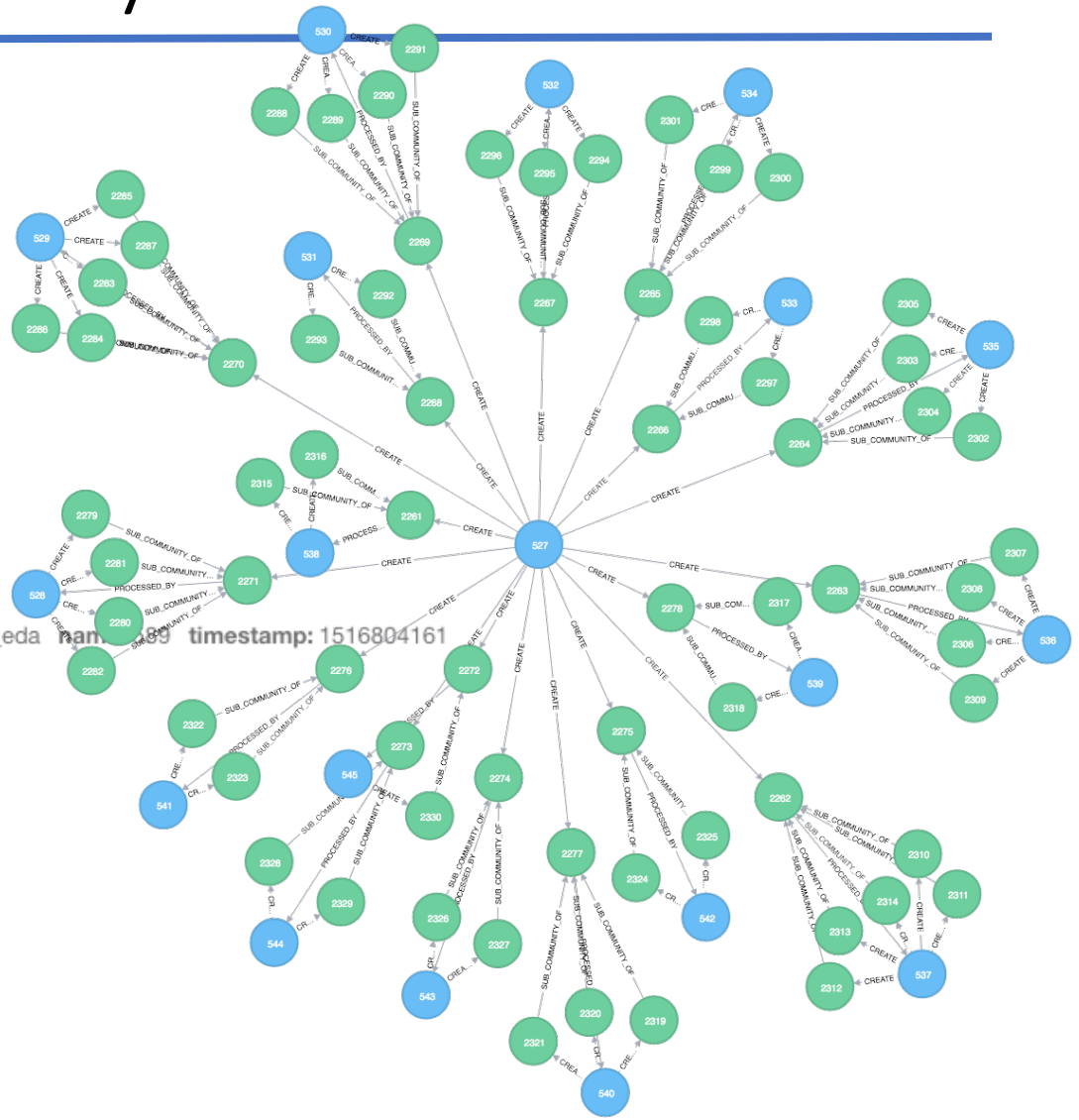
To compare results we implement an saving method on the Neo4j.

Blue nodes are order

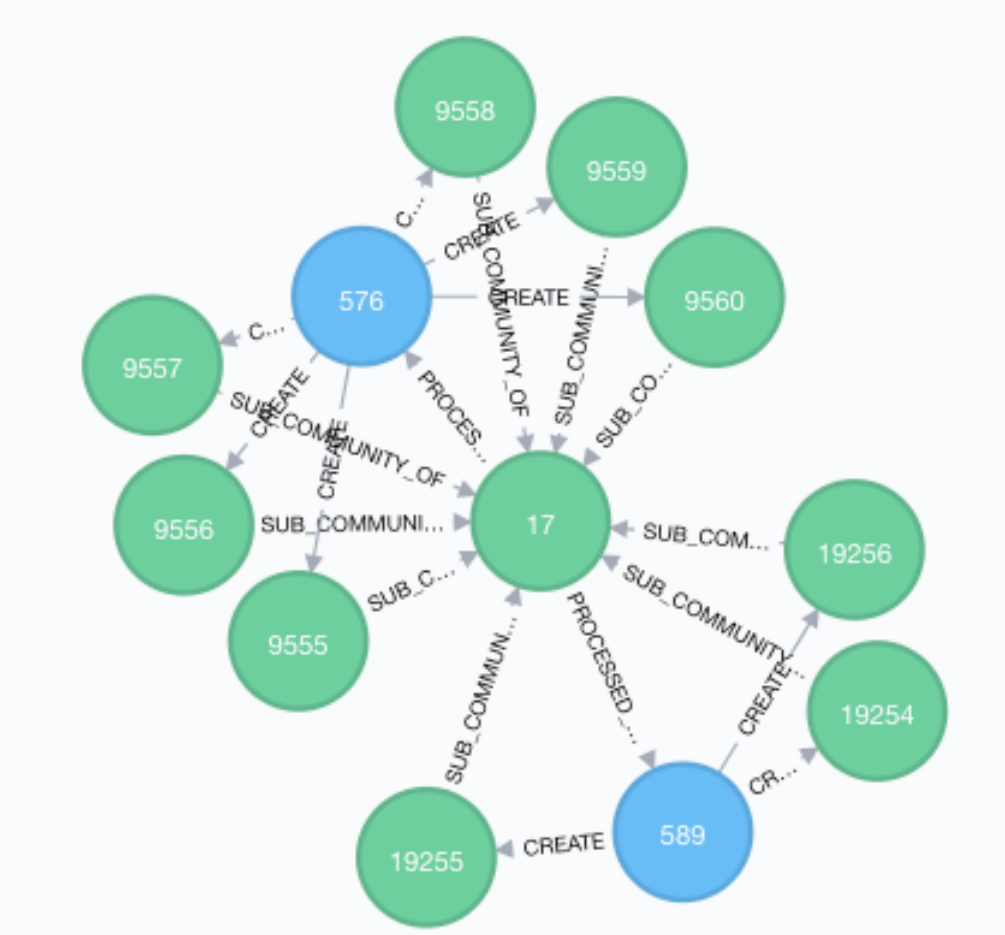
<id>: 110642 algorithm: spectralClustering_c:3_normalized_eda nam 589 timestamp: 1516804161

Green nodes Communities



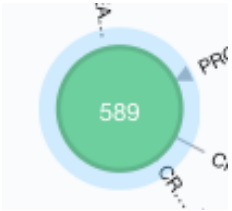


Order – Community Relation

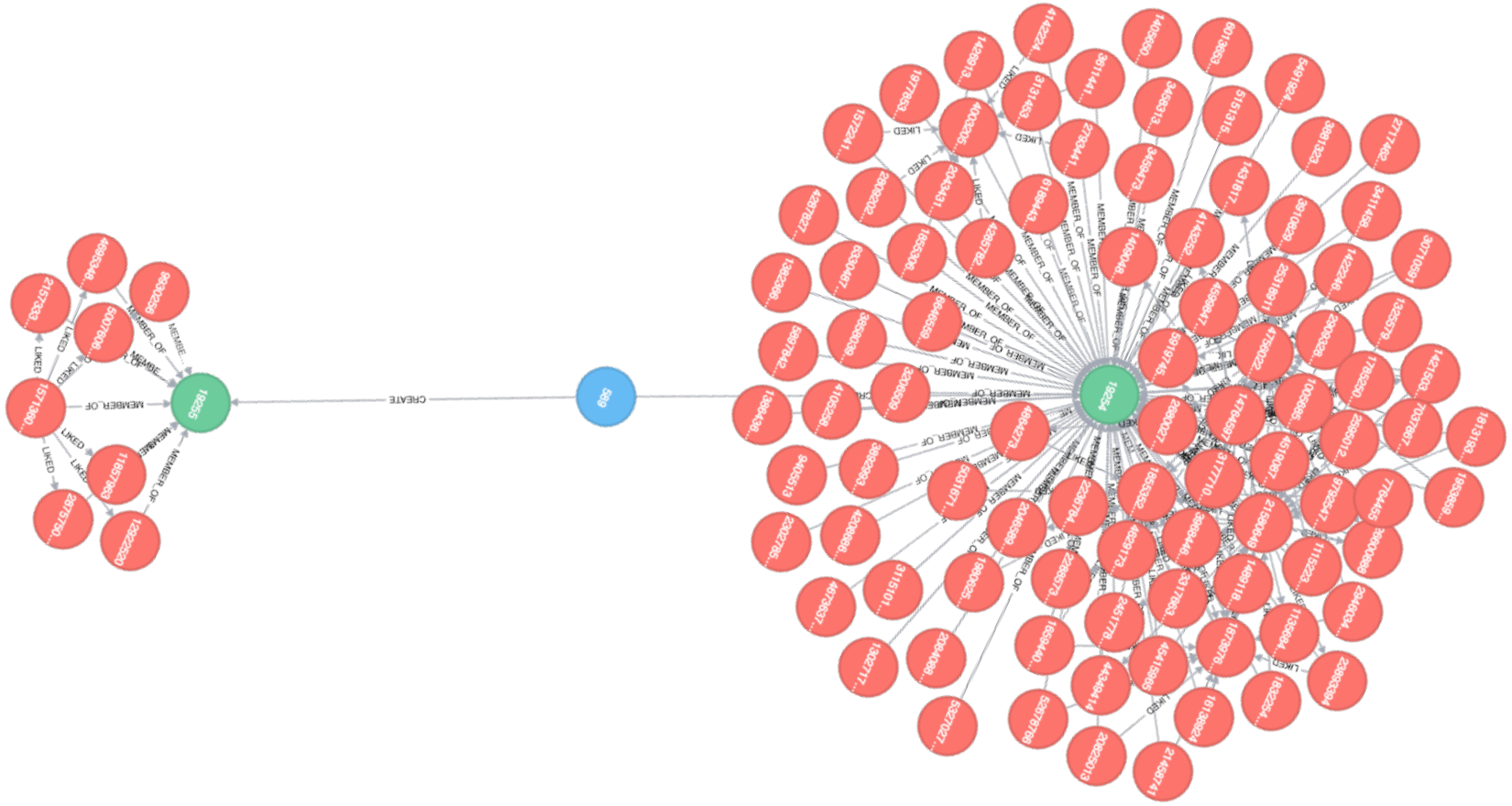


Blue nodes are order

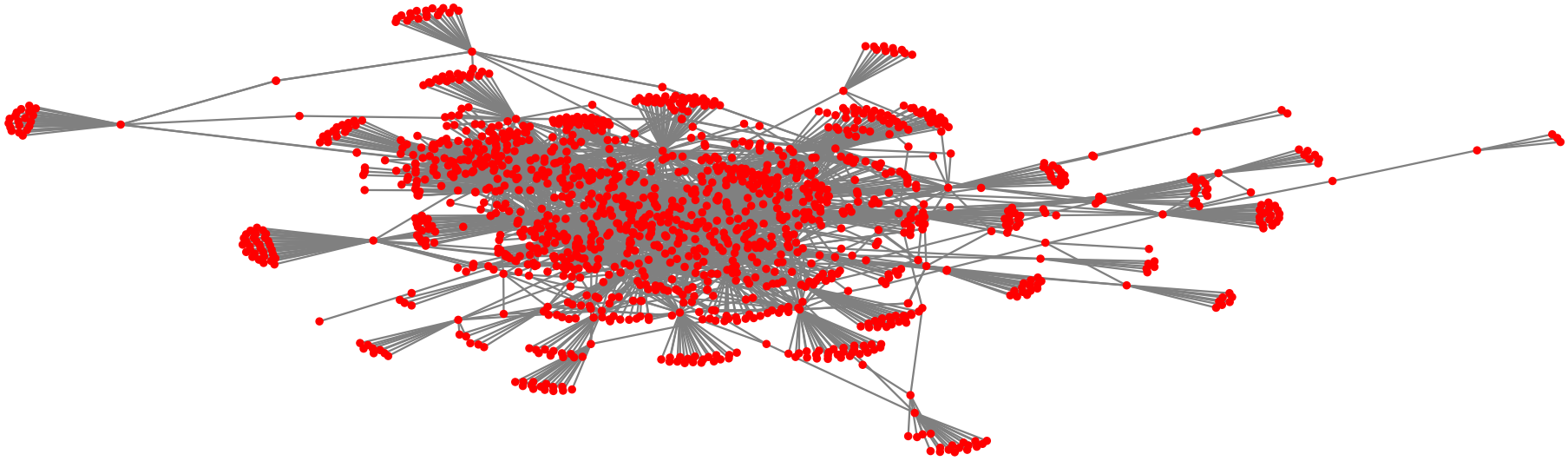
Green nodes Communities



Community – user Relation

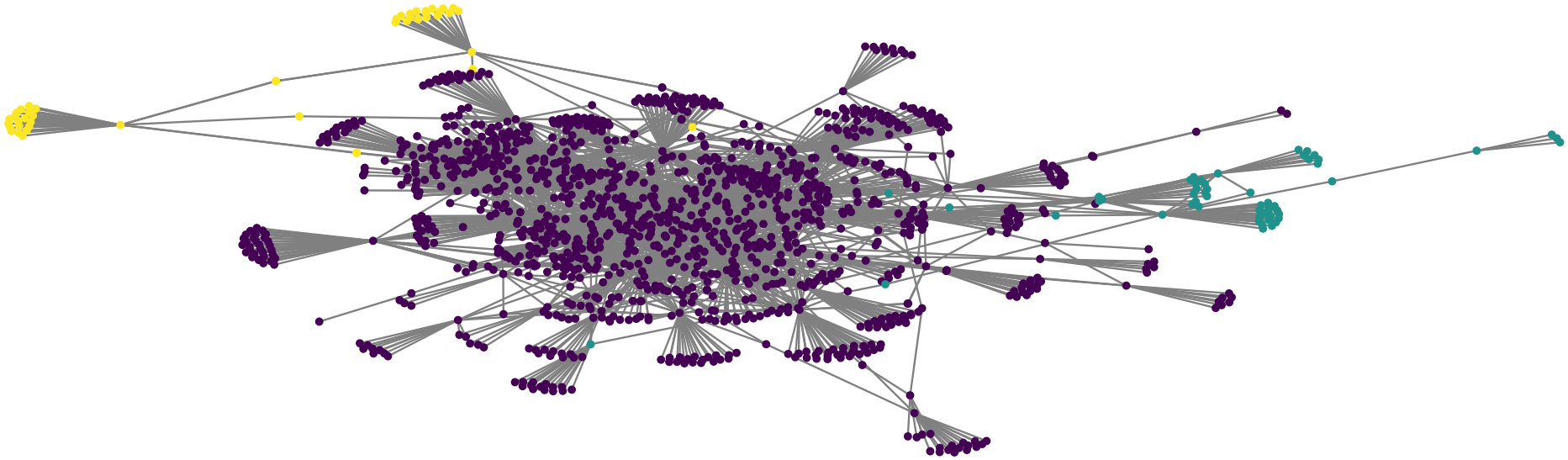


Spectral Clustering and Labeling



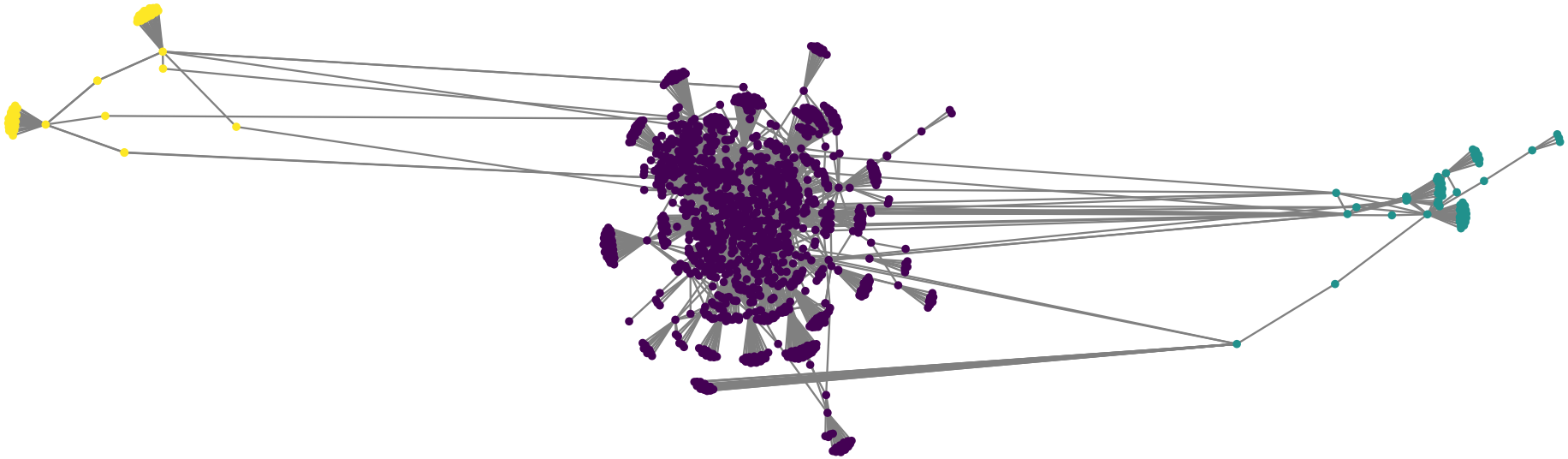
- 1357 nodes
- 2858 edges

Spectral Clustering and Labeling



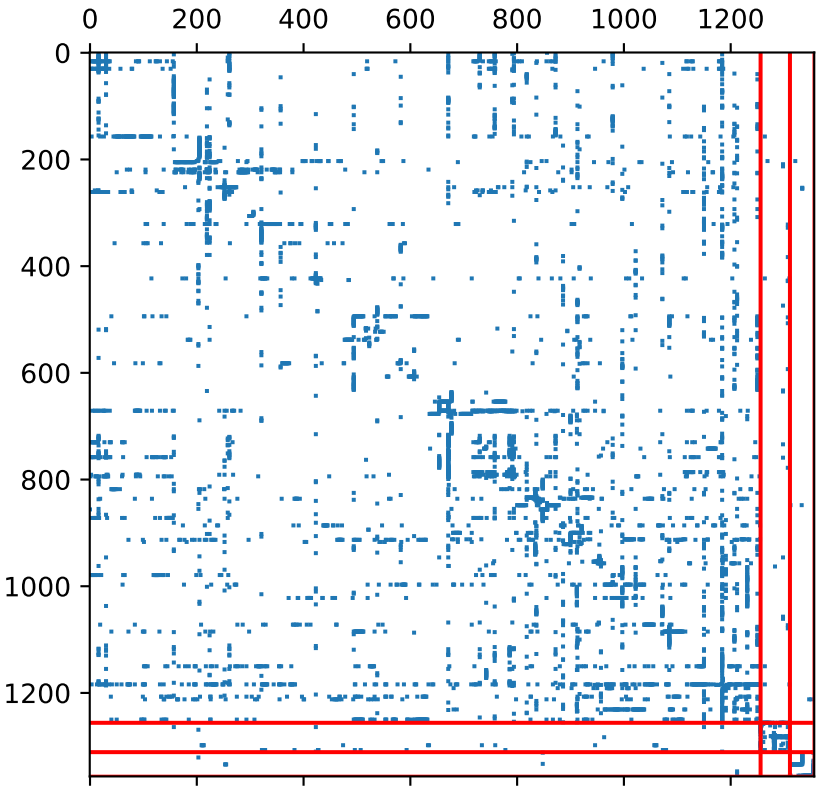
3 clusters

Spectral Clustering and Labeling



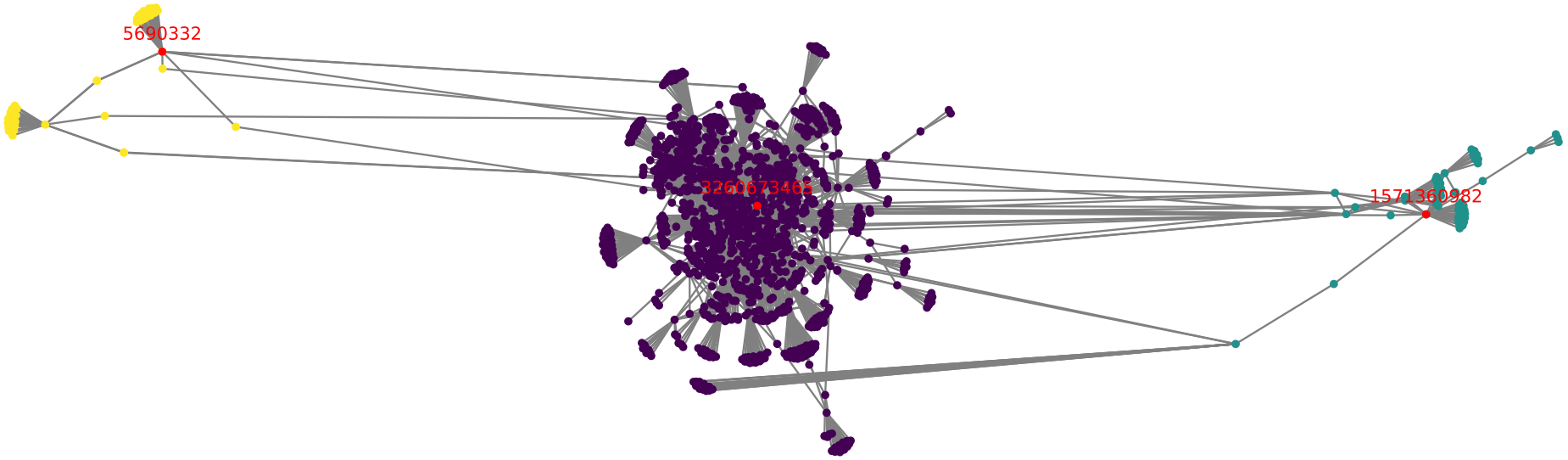
Separating the communities

Spectral Clustering and Labeling



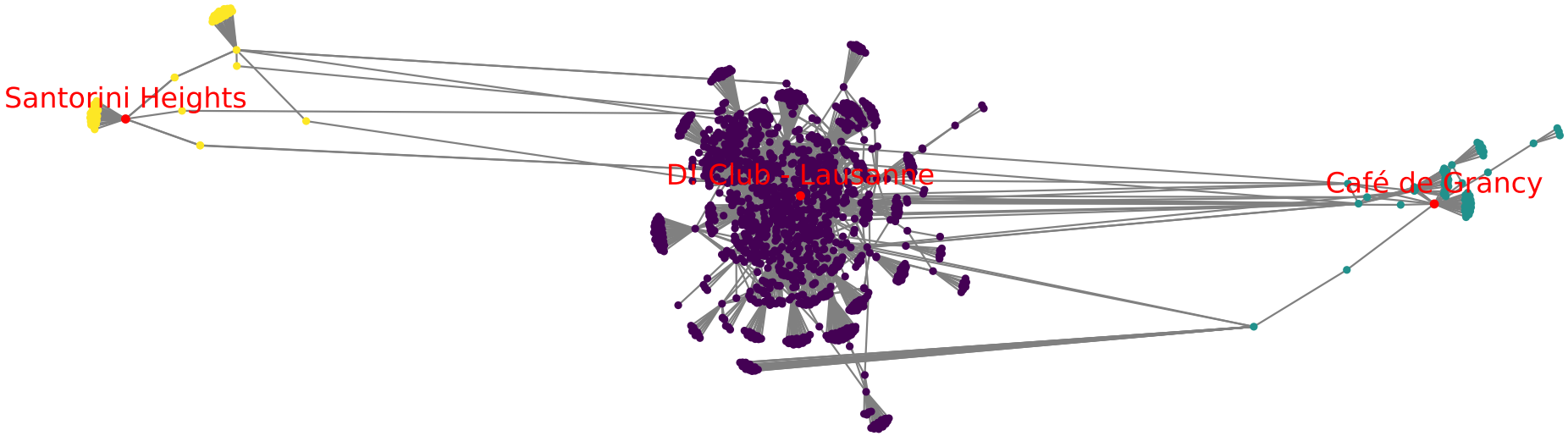
Ordered Adjacency Matrix

Spectral Clustering and Labeling



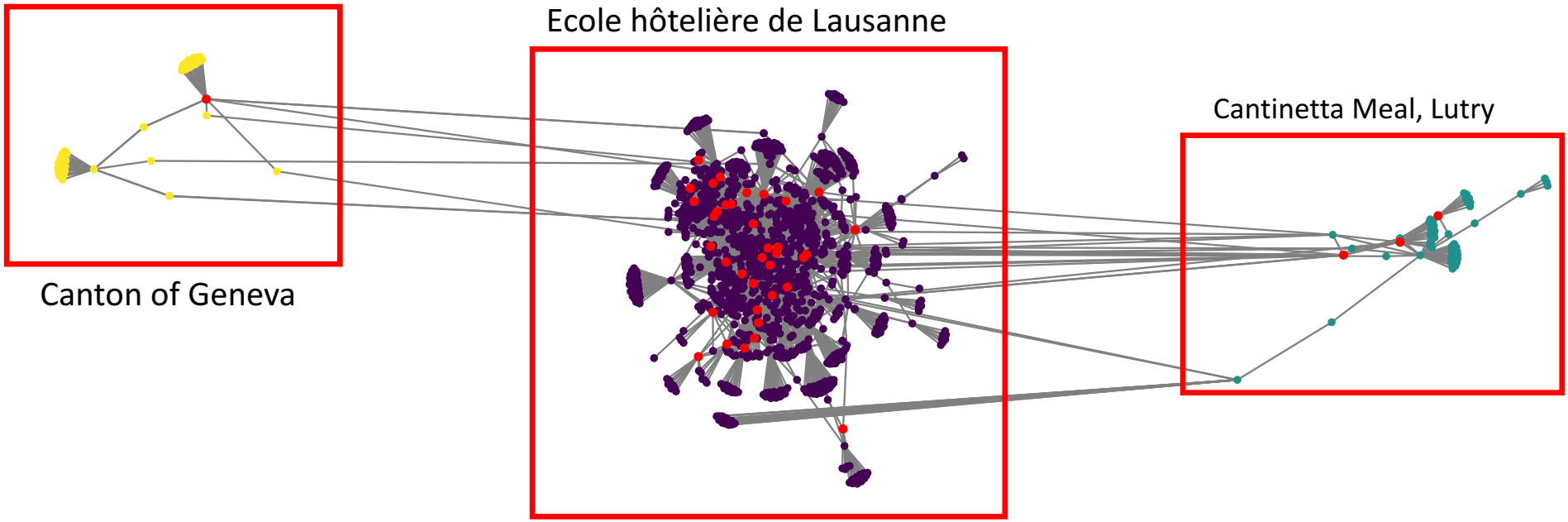
Hub users in each community

Spectral Clustering and Labeling



Last visited place by the hub users

Spectral Clustering and Labeling



Common place visited by the users

Comparison with Other Methods

- Girvan-Newman Algorithm
  - Based on removal of edges with largest centrality
- Louvain Algorithm
  - Maximizing modularity
- Stochastic Block Model
  - Determining typical pattern i.e. when nodes are connected mostly to other nodes of the same group.

Comparison with Other Methods

- Similarity measure between algorithms
  - Algorithm 1 = { (a,b), (c,d,e), (f,g,h) }
  - Algorithm 2 = { (a,b,c), (d,e,f), (g,h) }
  - Cartesian Product of each cluster into a bucket
    - (a,b), (c,d), (c,e), (d,e), (f,g), (f,h), (g,h)
    - (a,b), (a,c), (b,c), (d,e), (d,f), (e,f), (g,h)

$$similarity = \frac{n_{intersection}}{\sqrt{n_{\text{pairw comb Cr1}} \cdot n_{\text{pairw comb Cr2}}}}$$

Comparison with Other Methods

- Equalizing Number of Clusters
  - Different methods → Different number of clusters
  - Thinking of clusters as nodes
  - Merging the result with higher number of clusters until the number of clusters are equal

Comparison With Other Methods

Similarity Scores	Girvan-Newman	Stochastic BM	Spectral Clustering	Louvain
Girvan-Newman	1	0.583	0.941	0.347
Stochastic BM	0.583	1	0.595	-
Spectral Clustering	0.941	0.595	1	0.335
Louvain	0.347	-	0.335	1

Conclusion

- Big Data → more qualified labels