Università degli studi ROMA TRE
Dipartimento di Ingegneria
Corso di Laurea Magistrale in Ingegneria Informatica

Search and Validation of Communities using semantic distance between users in a social network

Laureando
Giorgio Maria De Rango
440847

Relatore
Prof. Franco Milicchio

Correlatore
Ing. Roberto De Virgilio

Anno accademico 2016/2017

Community Detection

- **A community is a group of users who interact on common interests**
- **Community detection want to divide a social network in communities**

- **social networks are rich of informations and communities can be a great tool to aggregate and understand them**

- **many human behavior are much significant and predictable if you look at those from a group point of view. Thus community are useful in data analysis and recommendation systems.**

ROMA TRE
UNIVERSITÀ DEGLI STUDI

Community detection:
comparison between two classical approaches

features	semantic	topological
high topological density	✗	✓
high semantic density	✓	✗
domain based	✓	✓
noise resistant	✓	✗

Goals

- *Starting from groups of users that should represent the communities* "**cores**"
- *The task is to aggregate those "cores"* **through synergy between topological correlation and semantic proximity**
- *To find out* ***real*** *(topological correlation) and* ***realistic*** *(with semantic meaning)* ***communities***

ROMA TRE
UNIVERSITÀ DEGLI STUDI

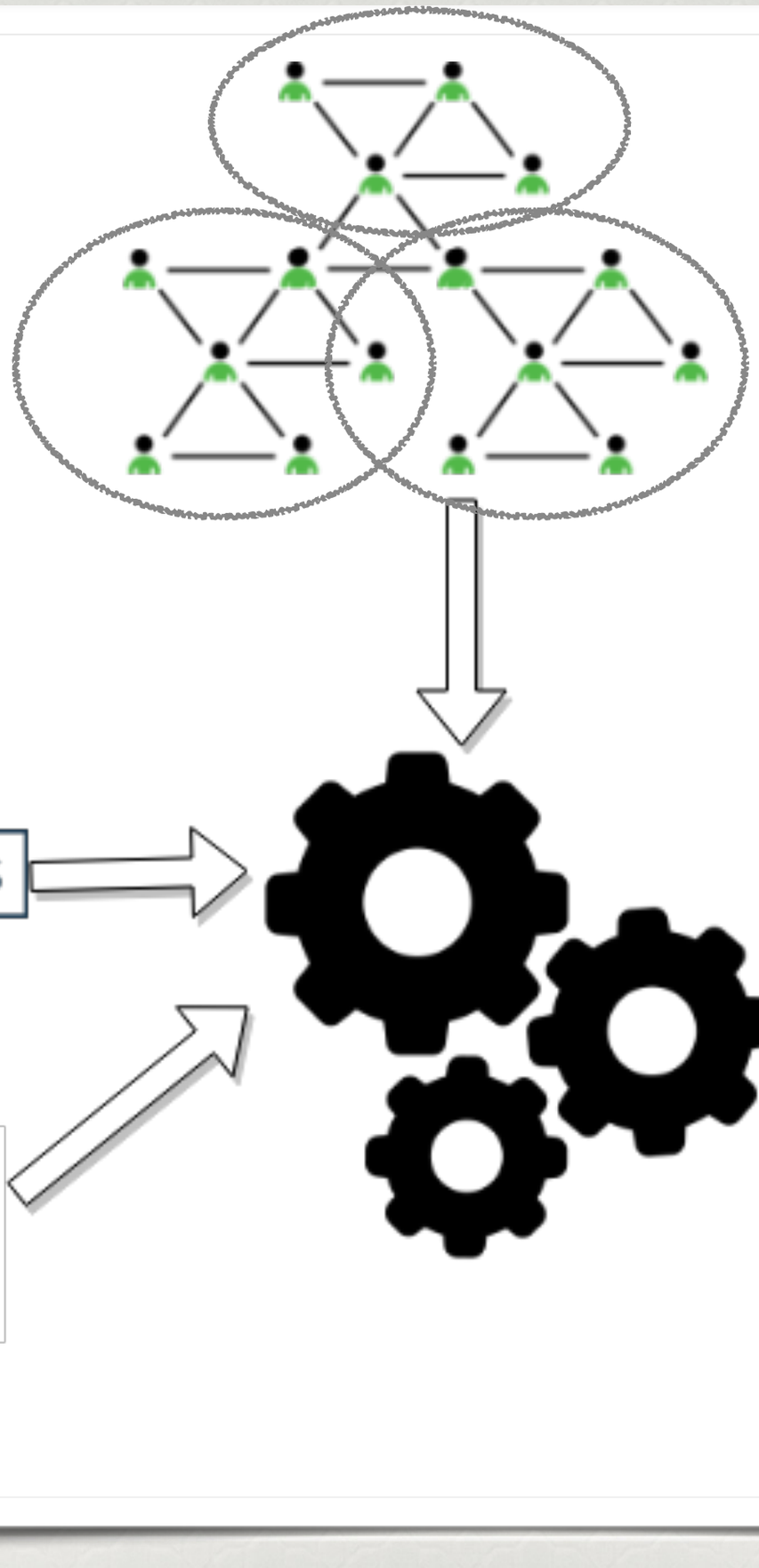
NetworkX

SDN应用路由算法实现工具之Networkx

node.js

www.sdnlab.com

kaggle

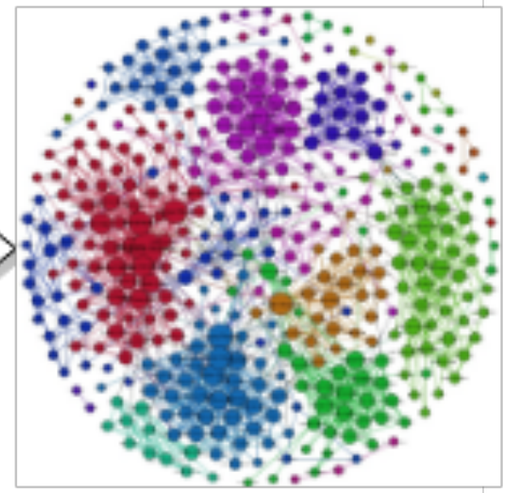


python™

mongoDB

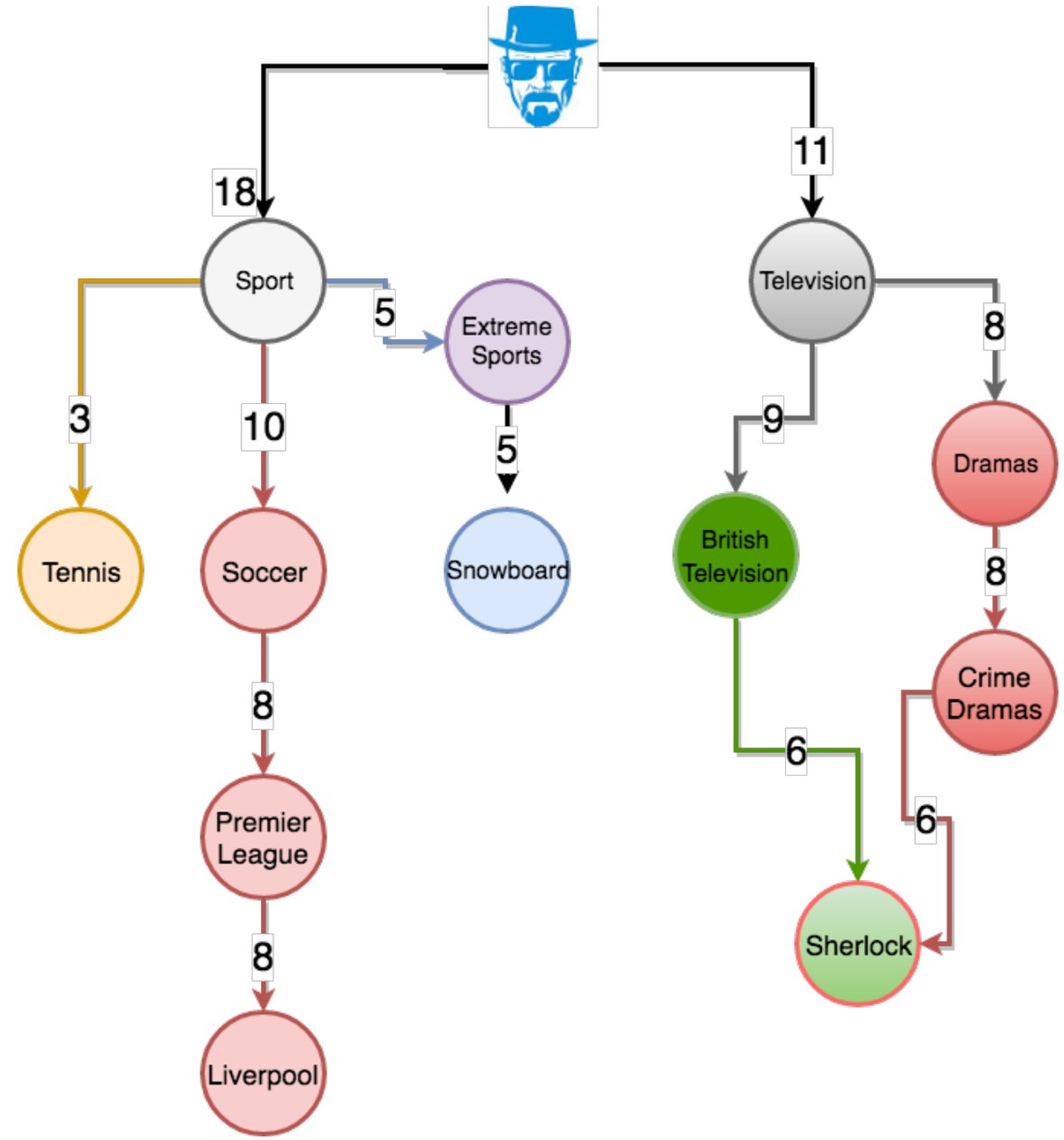
matplotlib

macromeasures



ROMA TRE
UNIVERSITÀ DEGLI STUDI

a user profile is a
weighted knowledge graph

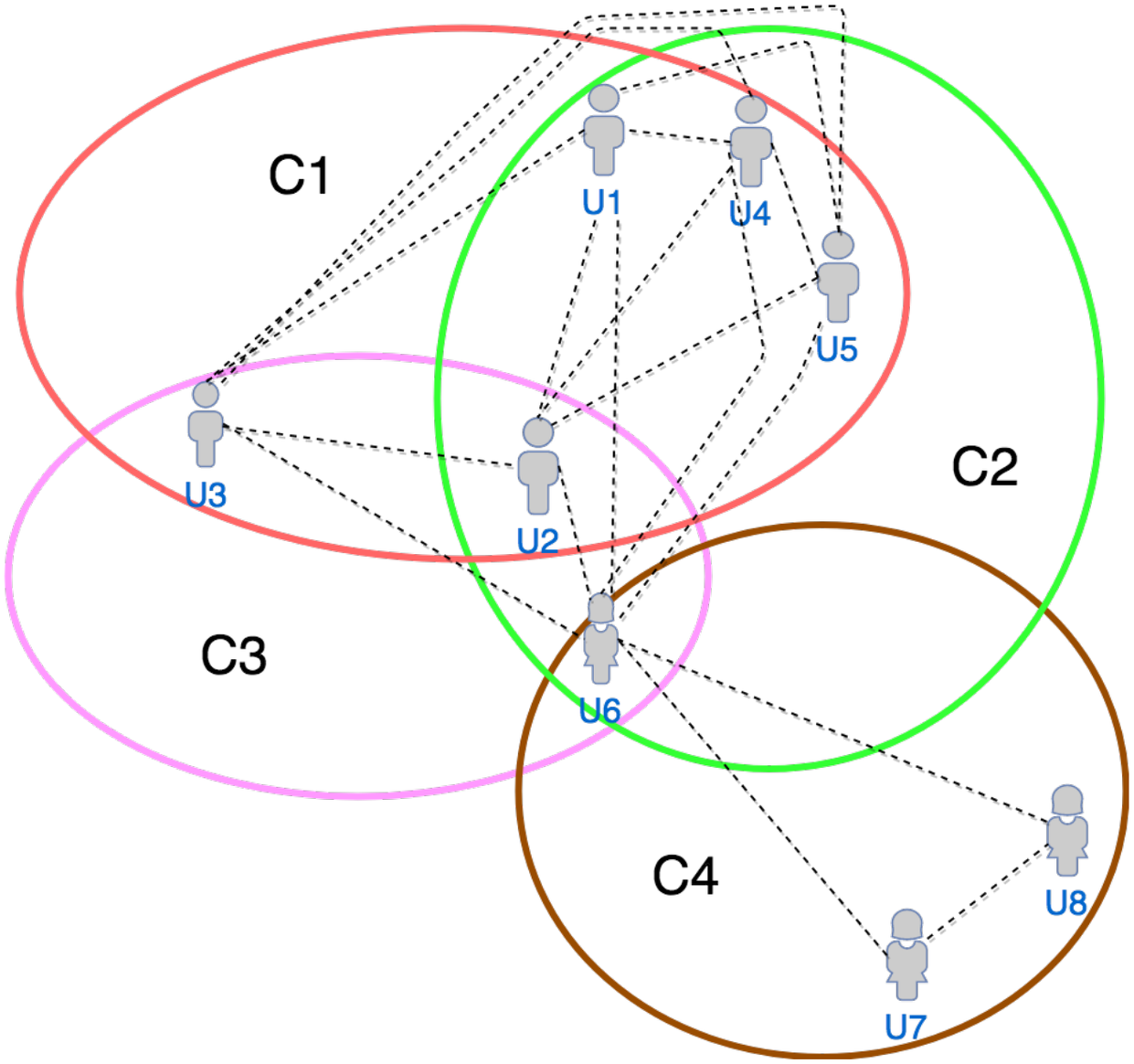


Interest	Score
Sport	18
Tennis	3
Soccer	10
Liverpool	8
Snowboard	5
Television	11

ROMA TRE
UNIVERSITÀ DEGLI STUDI

chosen community cores

Network → Clique



Users interests

macromeasures

Users profiles

ROMA TRE
UNIVERSITÀ DEGLI STUDI

user profiles

each interest has a score.

Interest	Score
Automotive	12
Cars	14
Racing cars	15
F3	15
Television	19
X-factor	19

Interest	Score
Automotive	29
Cars	29
Racing cars	29
F1	28
Ferrari	25
Vettel	23

Interest	Score
Automotive	40
Cars	40
Racing cars	38
F1	38
McLaren	33
Hamilton	33

Interest	Score
Automotive	27
Cars	27.5
Racing cars	27.3
F1	22

core profile

relaxed intersection between user profiles.
On k profiles an interest should stay at least in k - j profiles, with $j < k$.

core cohesion

U1

Interest	Score
Automotive	12
Cars	14
Racing cars	15
F1	0

U2

Interest	Score
Automotive	29
Cars	29
Racing cars	29
F1	28

distance = 0.137

U2

Interest	Score
Automotive	29
Cars	29
Racing cars	29
F1	28

U3

Interest	Score
Automotive	40
Cars	40
Racing cars	38
F1	38

distance = 0.0003

U3

Interest	Score
Automotiv	40
Cars	40
Racing cars	38
F1	38

U1

Interest	Score
Automotive	12
Cars	14
Racing cars	15
F1	0

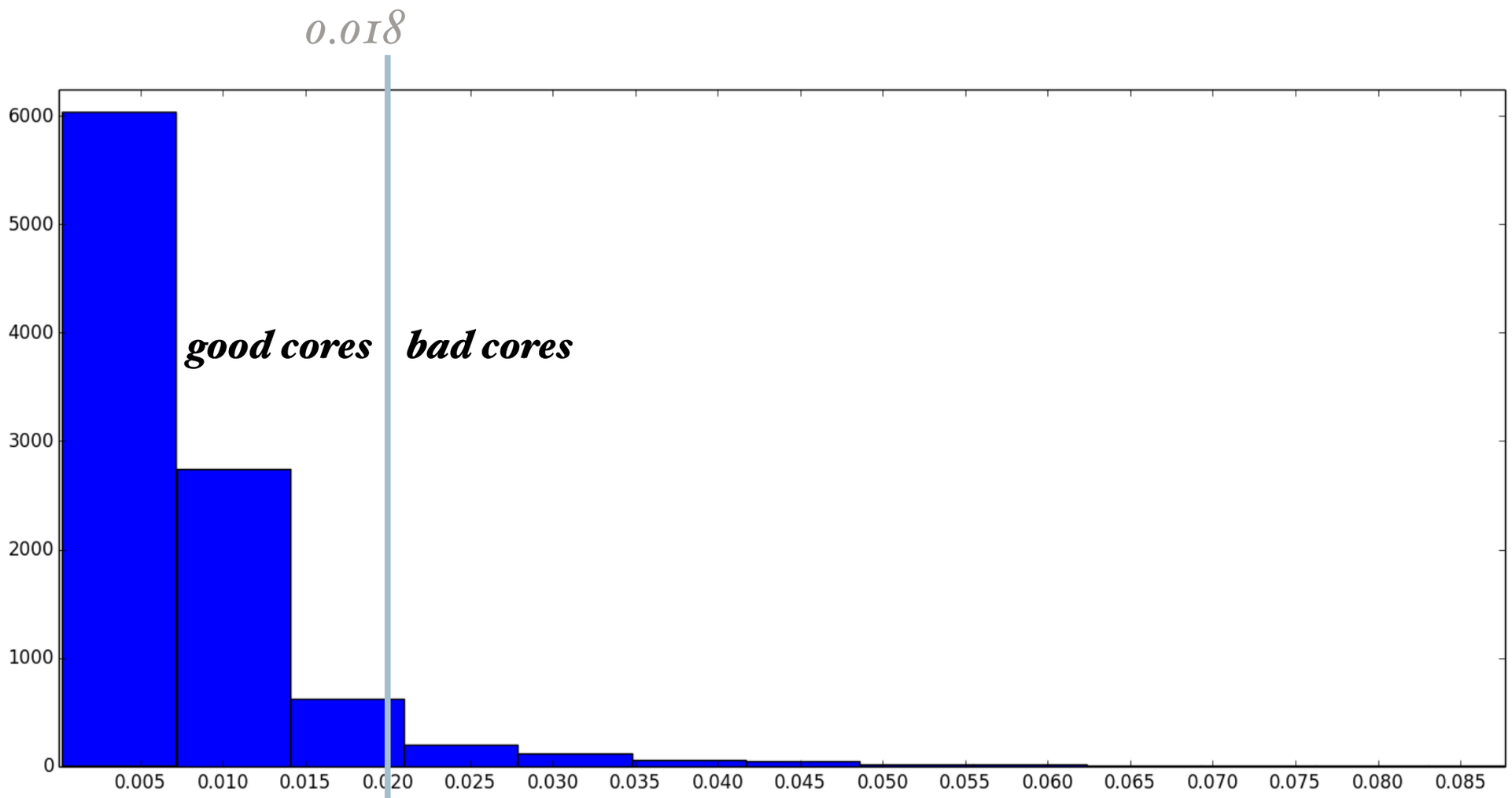
distance = 0.131

$$|\{u_1,u_2|u_1,u_2\in U\}|\sqrt{\prod_{i,j\in U}(1-cosine(p_i,p_j))}=0.017$$

Preprocessing

cohesion values distribution

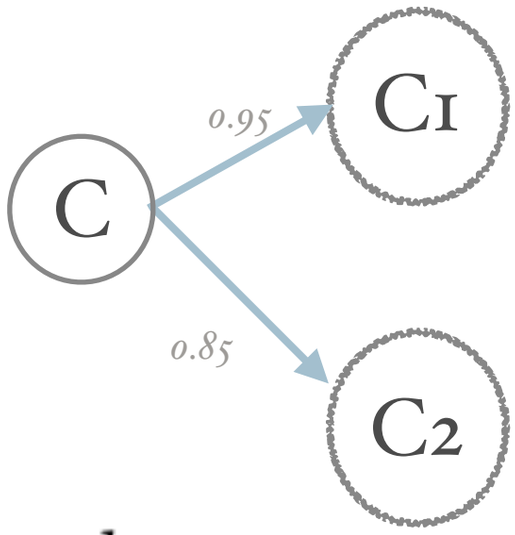
cohesion threshold



proximity graph

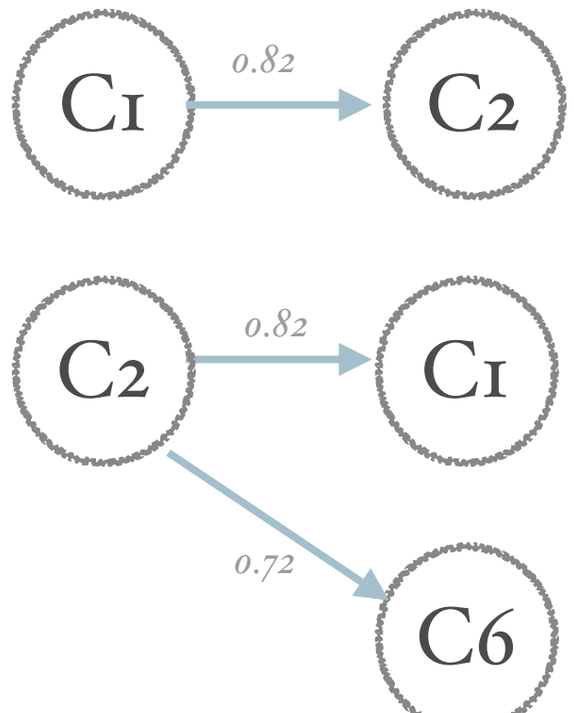
biggest core

c neighbors

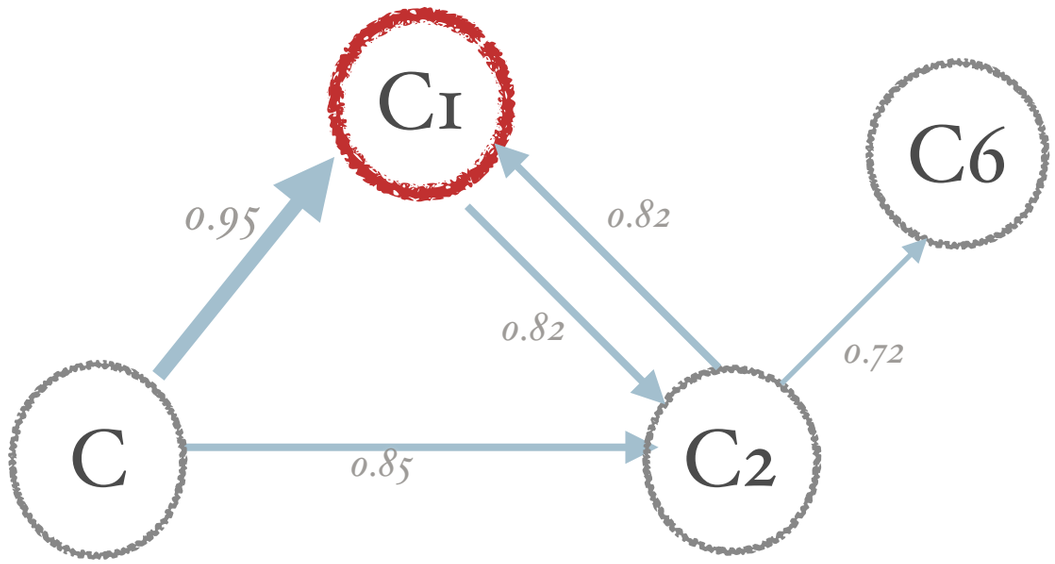


$k - \epsilon$ *shared users*

C1,C2 neighbors



=



C1 has many entering high weight arcs

cohesions of di C, C1,C2,C6 > threeshold

C, C1,C2 are labelled as visited nodes

Pagerank

+

not uniform probability destribution

on arcs

using semantic similarity between cores

and nodes

$$((1+\text{mean})^3)/\sqrt{\#\text{ of users}}$$

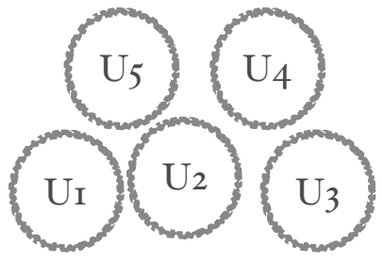
=

Attractor node

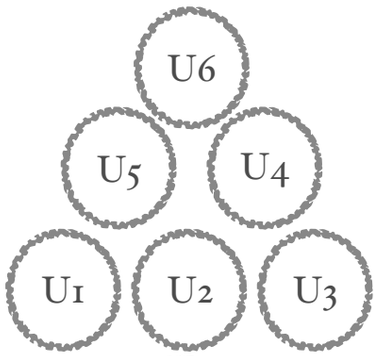
attractor node N

arc with highest weight

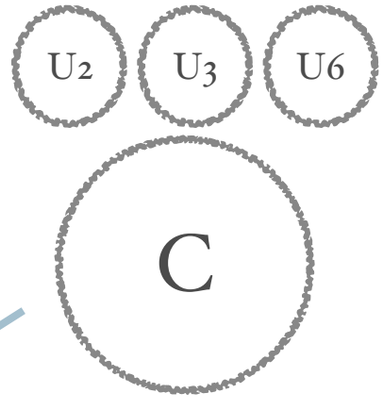
most similar node to N



C1



CC1

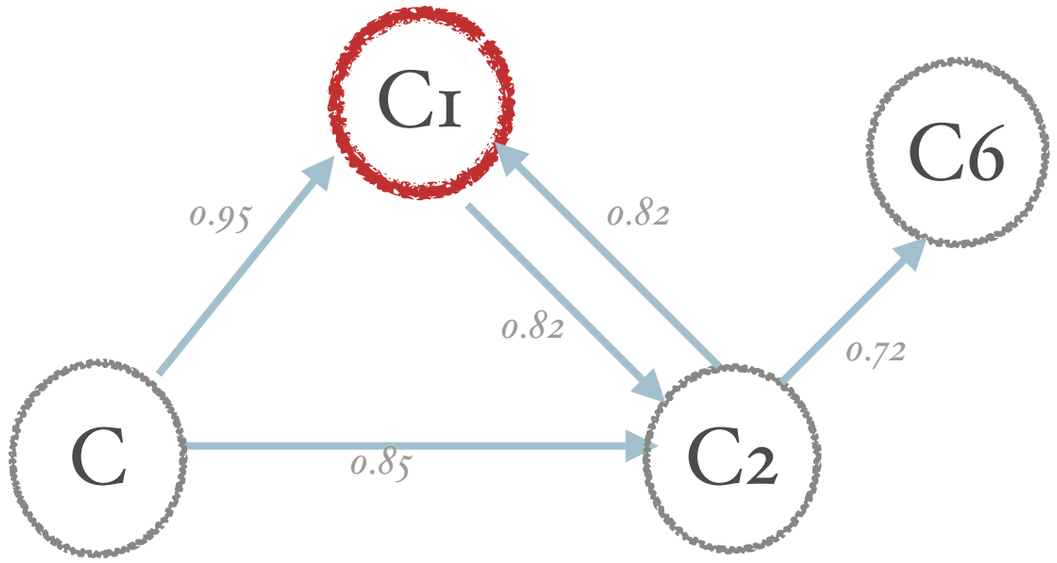


C

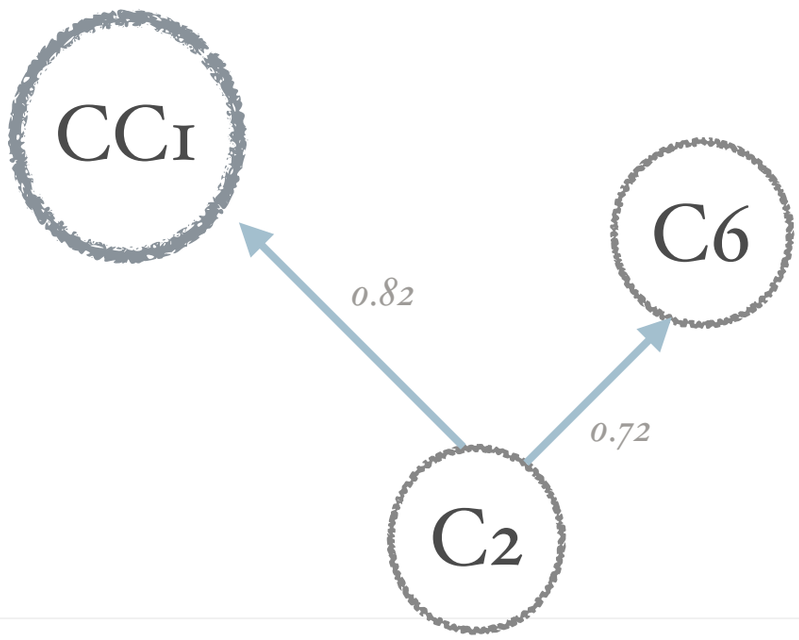
if |CC1' s profile| > t

fusion confirmed

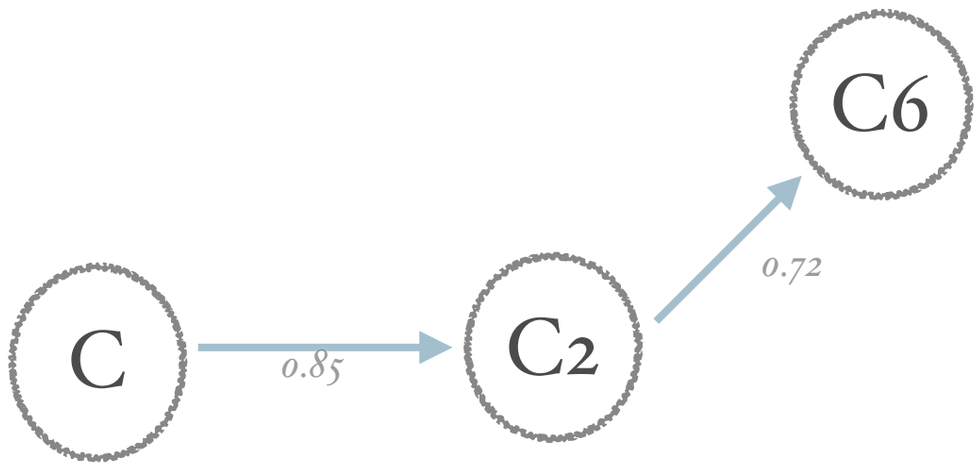
Updating proximity graph



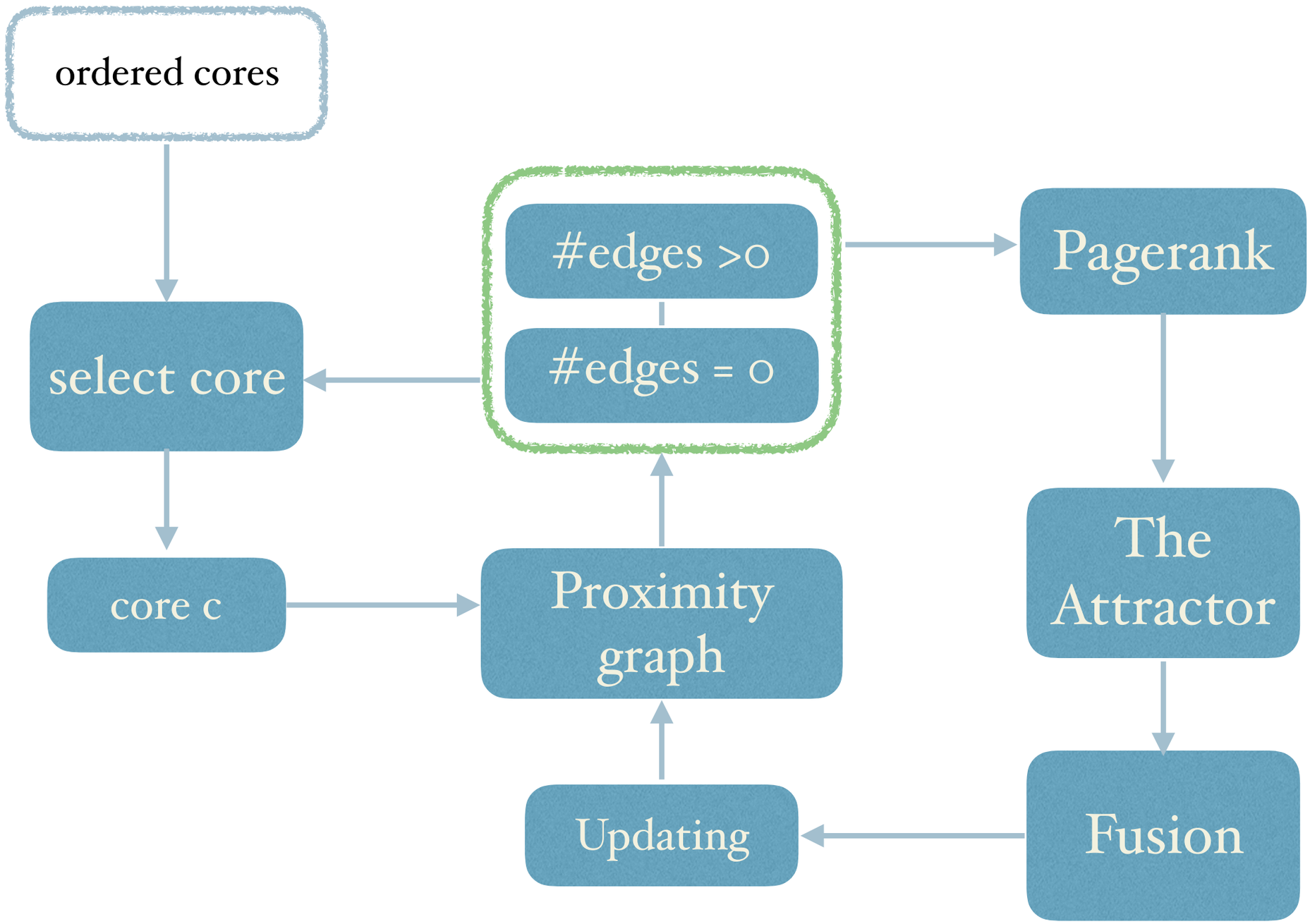
Approved



Not approved



Modello di un' iterazione

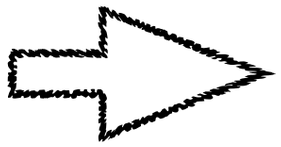


ROMA TRE
UNIVERSITÀ DEGLI STUDI

Dataset

kaggle

3500 users



36500 cliques

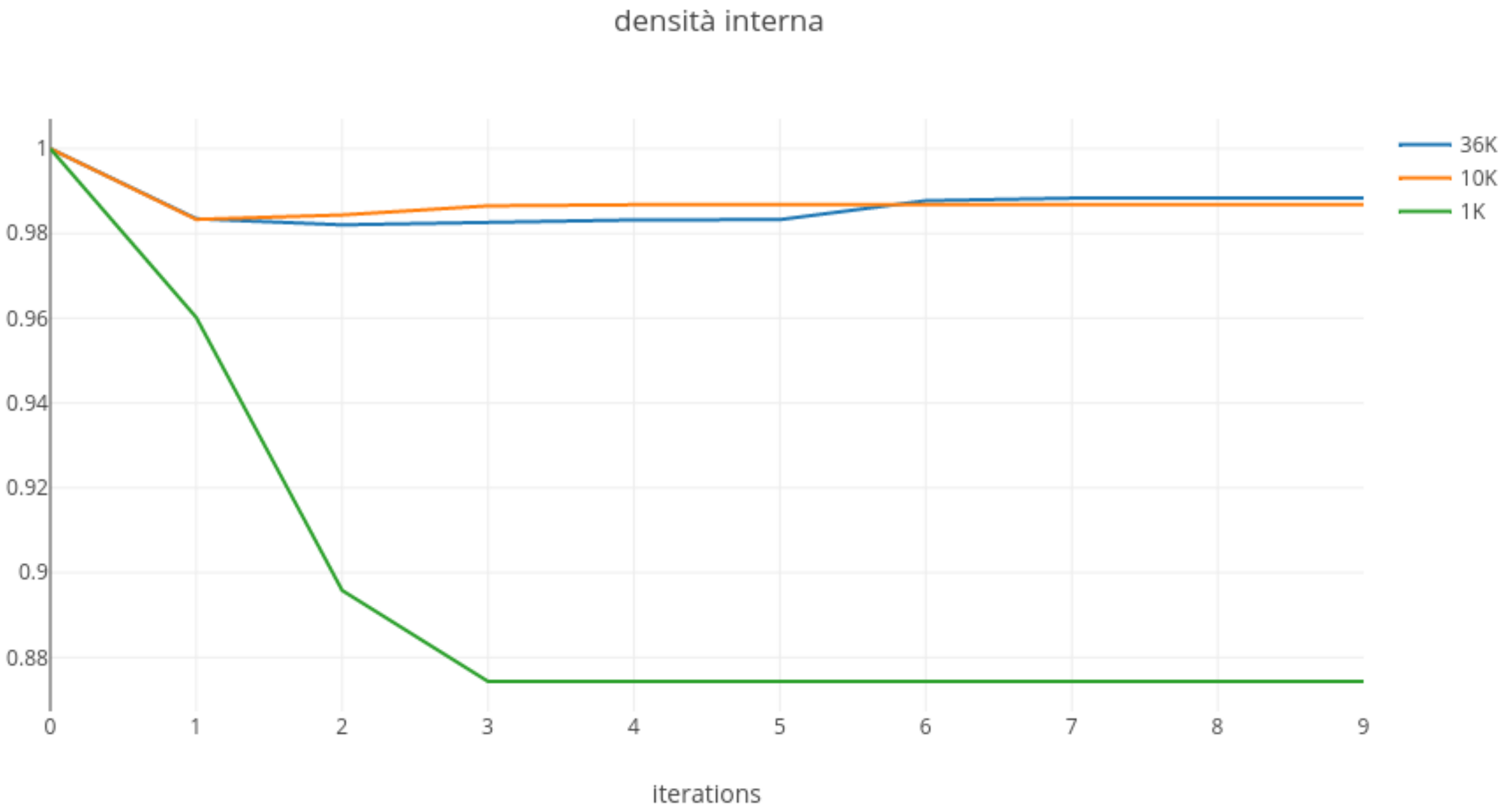
friends + followers > 200

80 interest per user

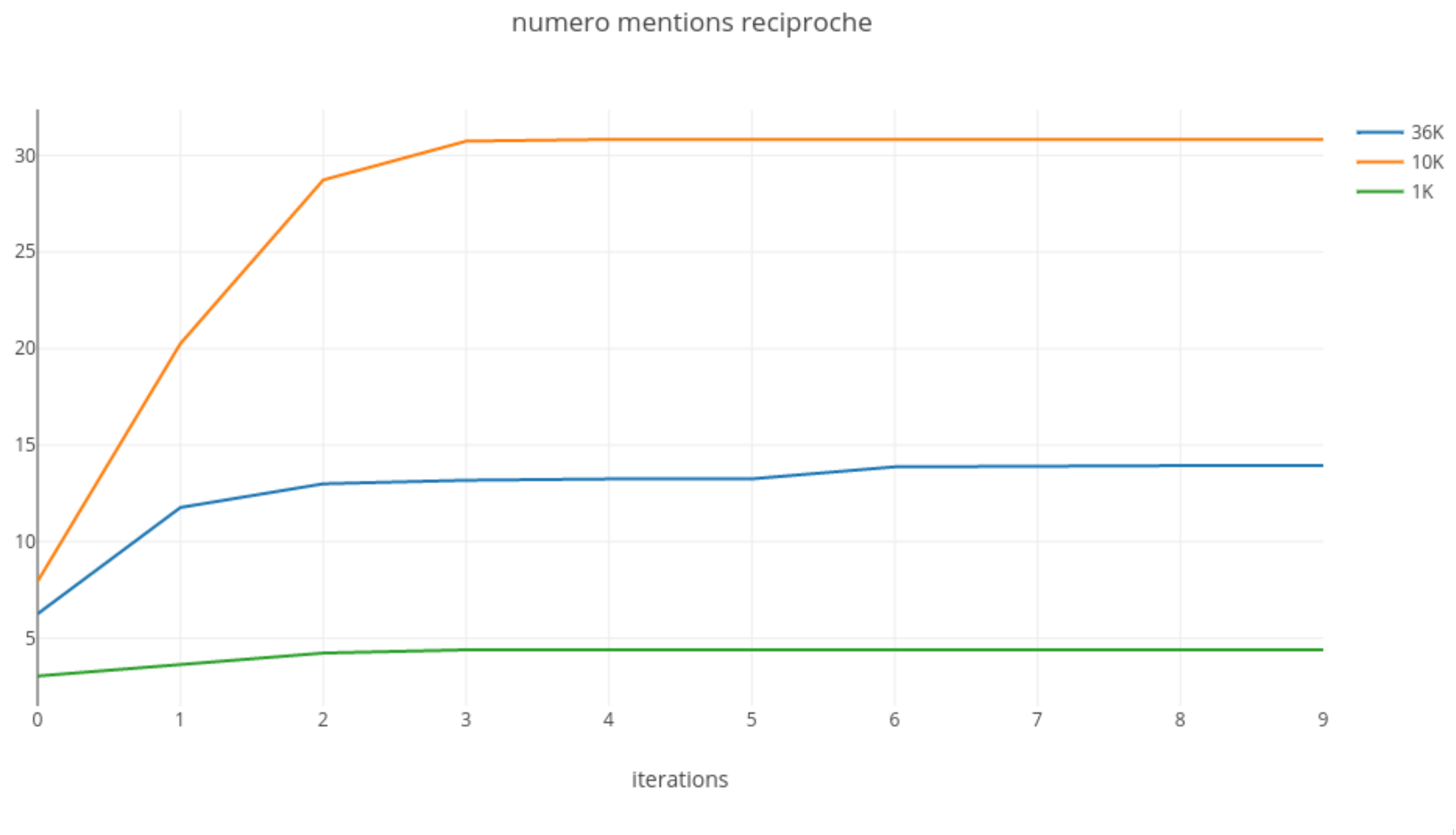
dense graph

98 clique per user

Effectiveness - internal density



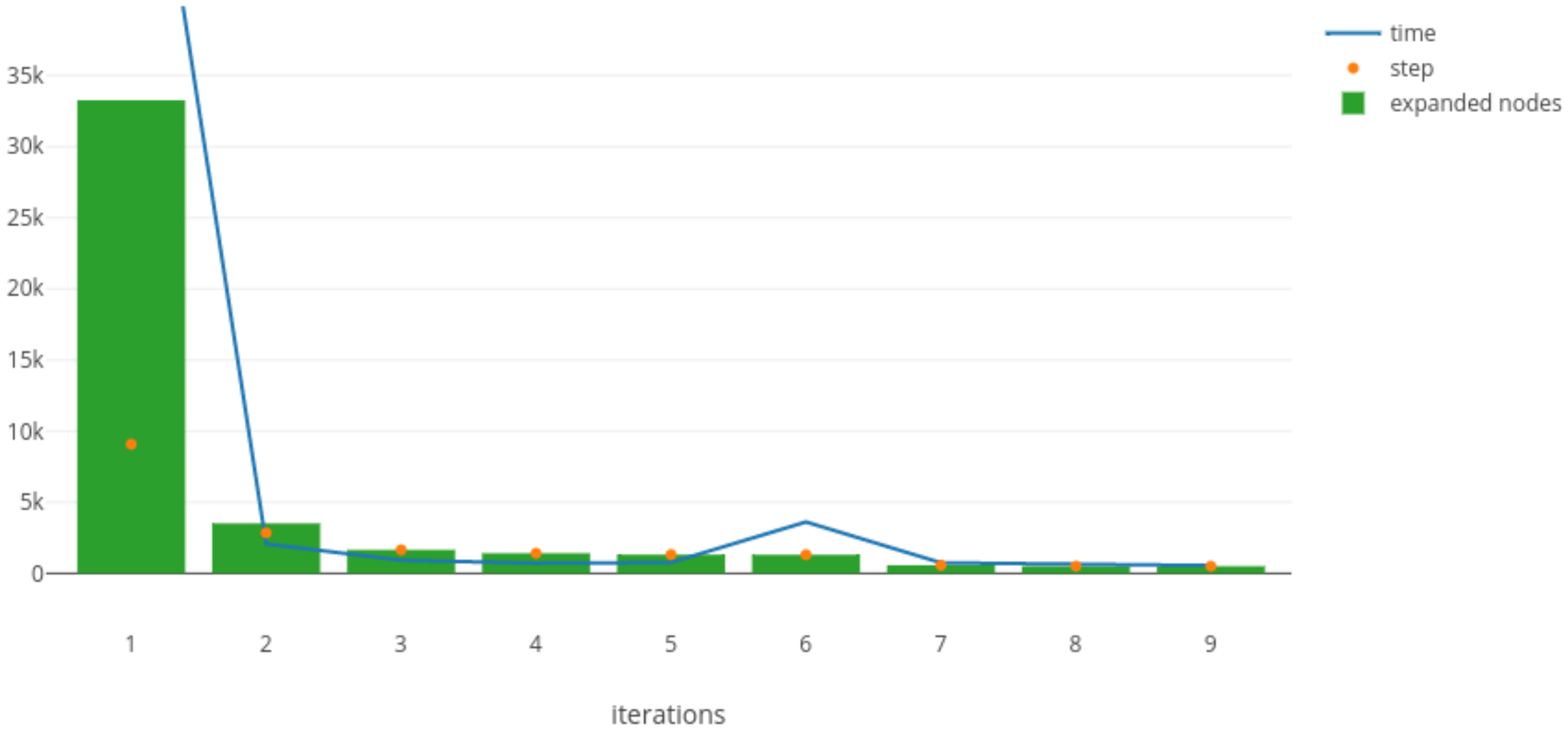
Effectiveness - mentions average in a community



Effectiveness - profile overlapping with its community profile

Effectiveness

from 36k core to 12 k communities
comparison between exec time and expanded nodes



Obtained results

A model and a framework for community detection

- *domain independent*
- *parametric*
- *noiseless*
- *effective*
- *scalable and with a distributed computation nature taking advantage of connected components independence*

Future works

- *analysis of results with different parameters values*
- *using a different topological approach (different from clique)*
- *test on different domains*
- *model patterns on obtained outputs*
- *community prediction*

The end ..