

*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 aims at dividing a social network into communities**
- **social networks are rich of information and communities can be a great tool to aggregate and understand them**
- **human behaviors are significant and predictable if you look at those from a group point of view. Thus communities are useful in data analysis and recommendation systems.**

Community detection: comparison between two classical approaches

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

Goal

Starting from groups of users that should represent the “cores” of the communities the task is

to aggregate those “cores” through the synergy between topological correlation and semantic proximity

to find out real (topological correlation) and realistic (with semantic meaning) communities

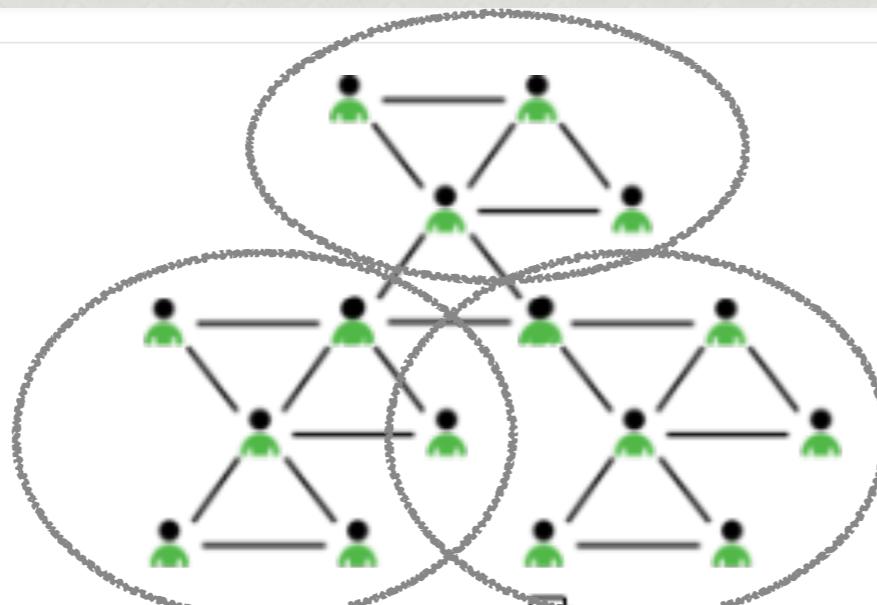


NetworkX

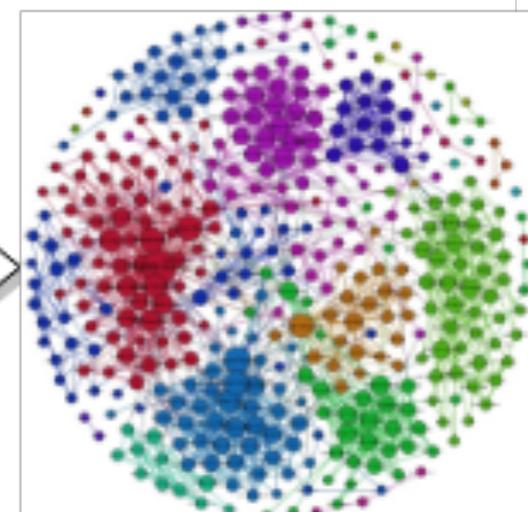
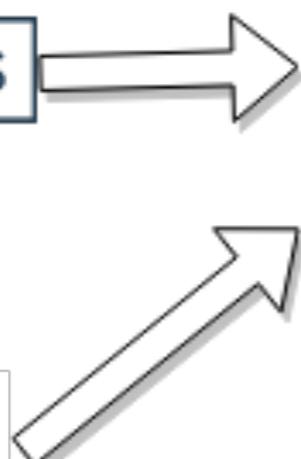
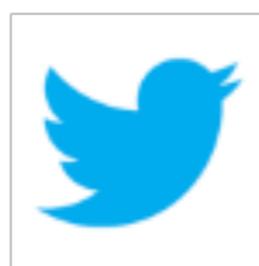
SDN应用路由算法实现工具之Networkx



kaggle™

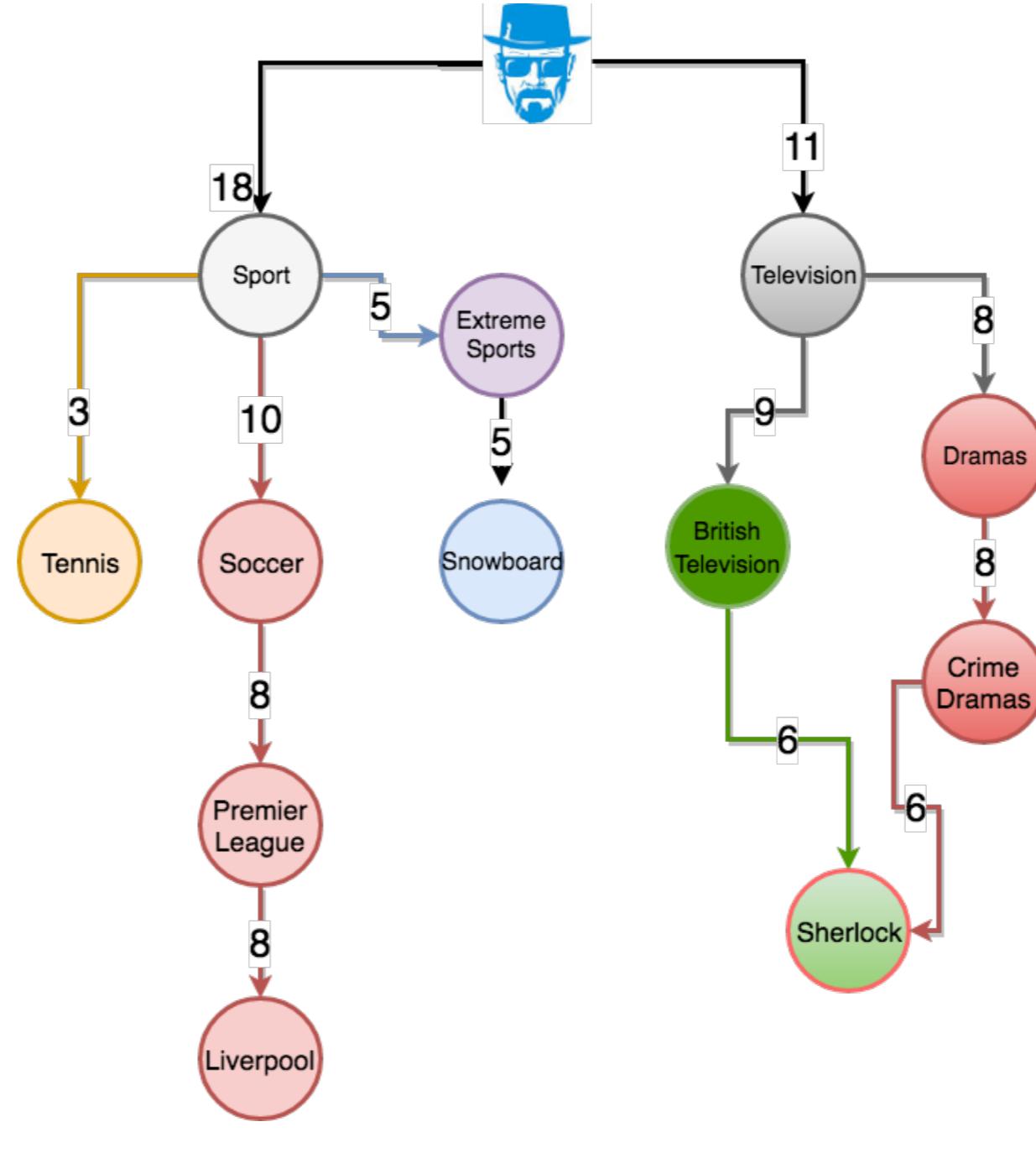


macromeasures

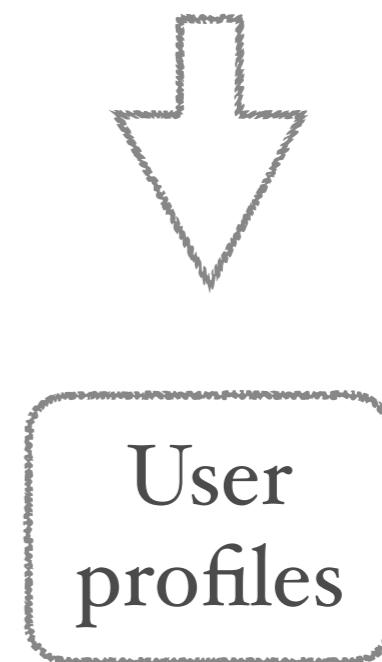
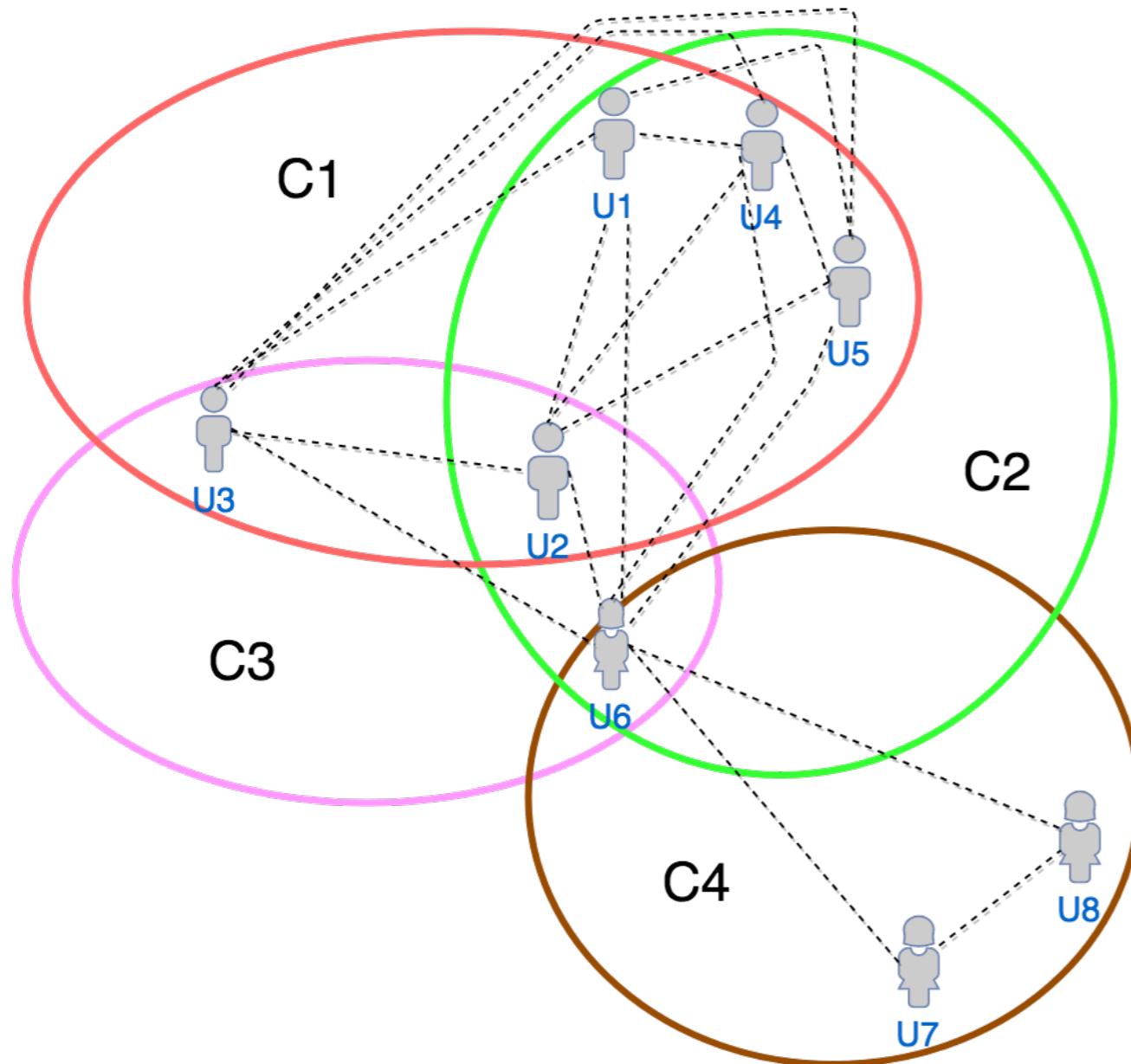
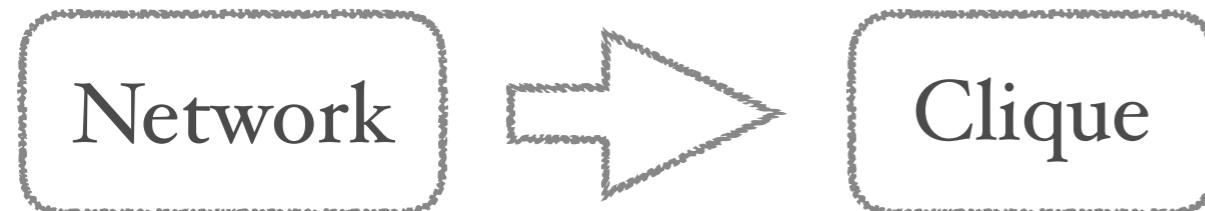




*a user profile is a
weighted knowledge graph*



chosen community cores



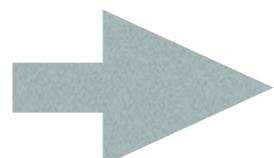
user
profiles

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

U_I	Interest	Score
Automotive	12	
Cars	14	
Racing cars	15	
FI	0	

U_2	Interest	Score
Automotive	29	
Cars	29	
Racing cars	29	
FI	28	

distance = 0.137

U_2	Interest	Score
Automotive	29	
Cars	29	
Racing cars	29	
FI	28	

distance = 0.0003

U_3	Interest	Score
Automotiv	40	
Cars	40	
Racing cars	38	
FI	38	

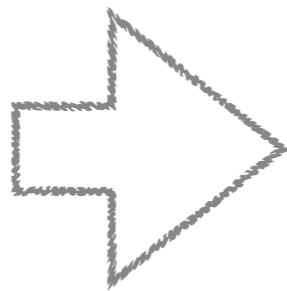
U_I	Interest	Score
Automotive	12	
Cars	14	
Racing cars	15	
FI	0	

distance = 0.131

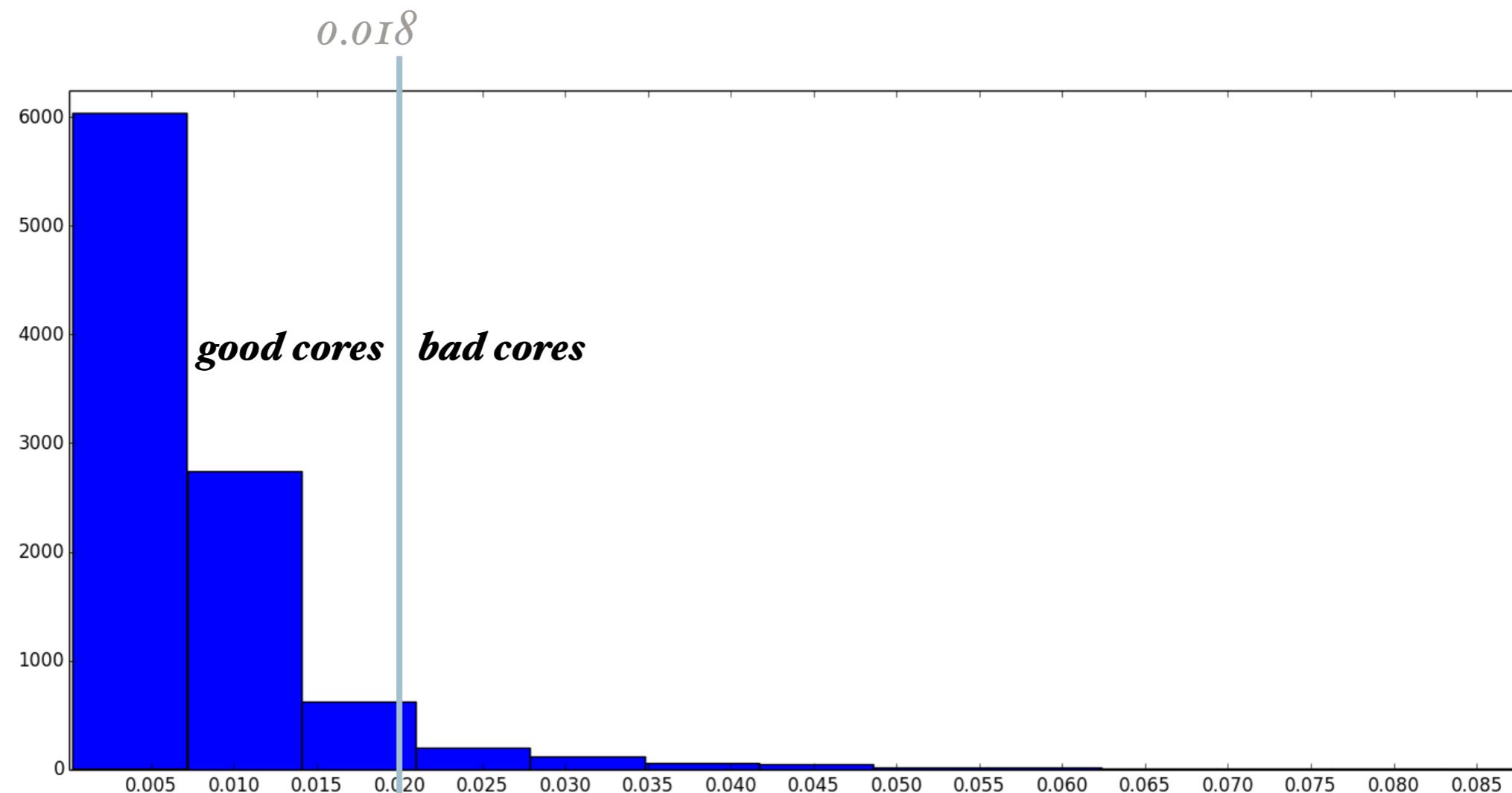
$$\sqrt{\prod_{i,j \in U} (1 - \text{cosine}(p_i, p_j))} = 0.017$$

Preprocessing

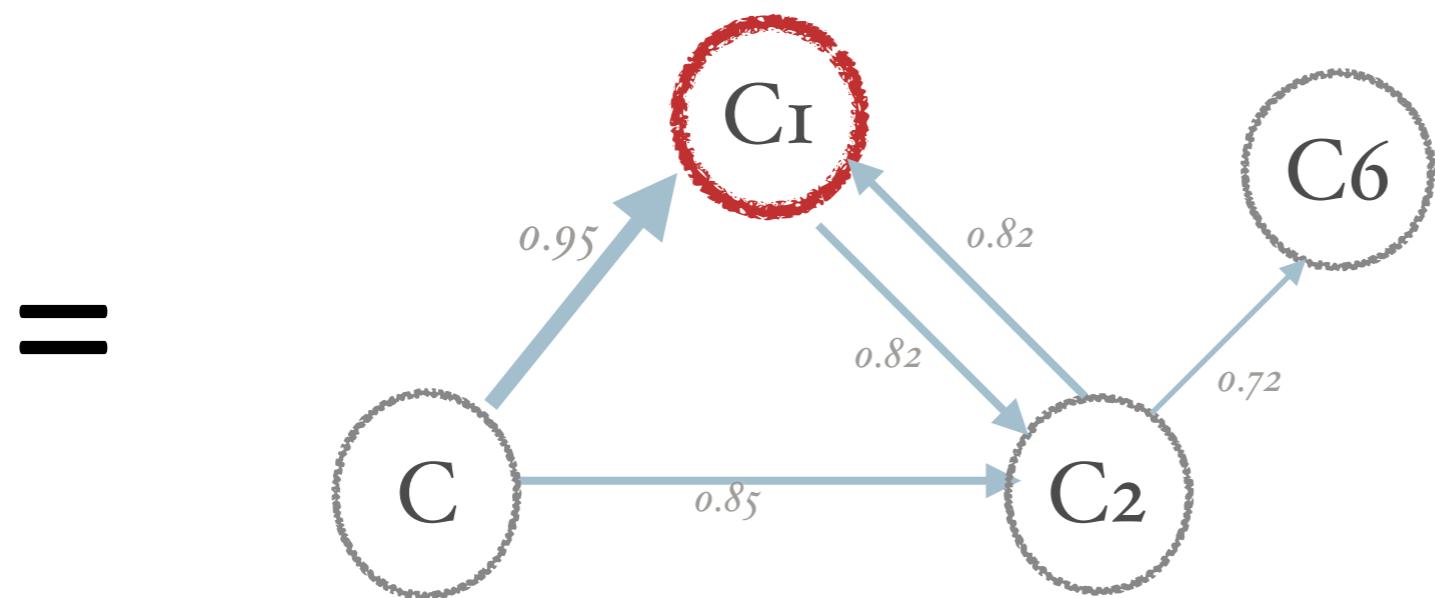
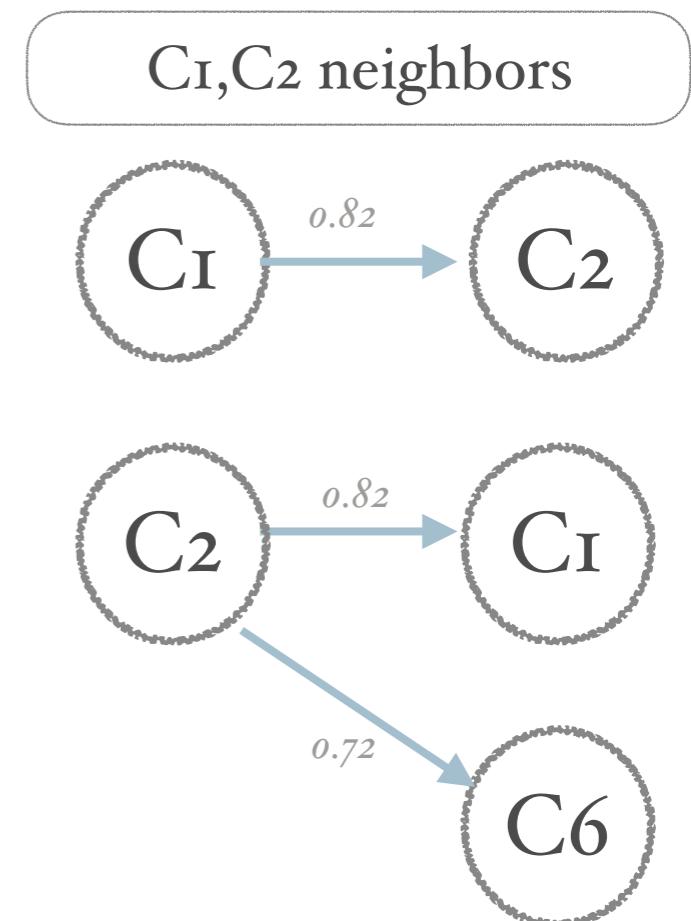
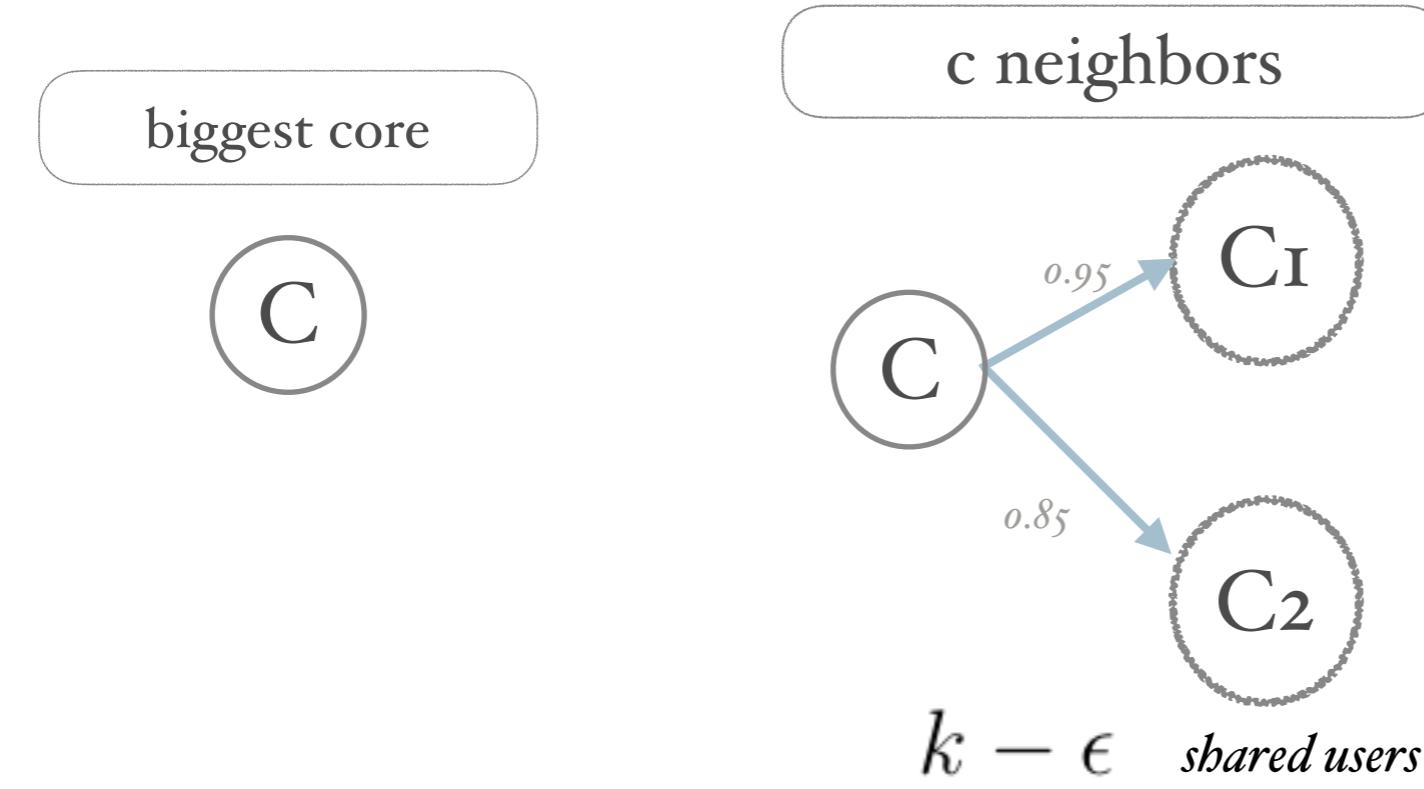
cohesion value
distribution



cohesion
threshold



proximity graph



C_1 has many entering high weight arcs

cohesions of $di(C, C_1, C_2, C_6) >$ threshold

C, C_1, C_2 are labelled as visited nodes

Pagerank

+

not uniform probability
distribution

on arcs

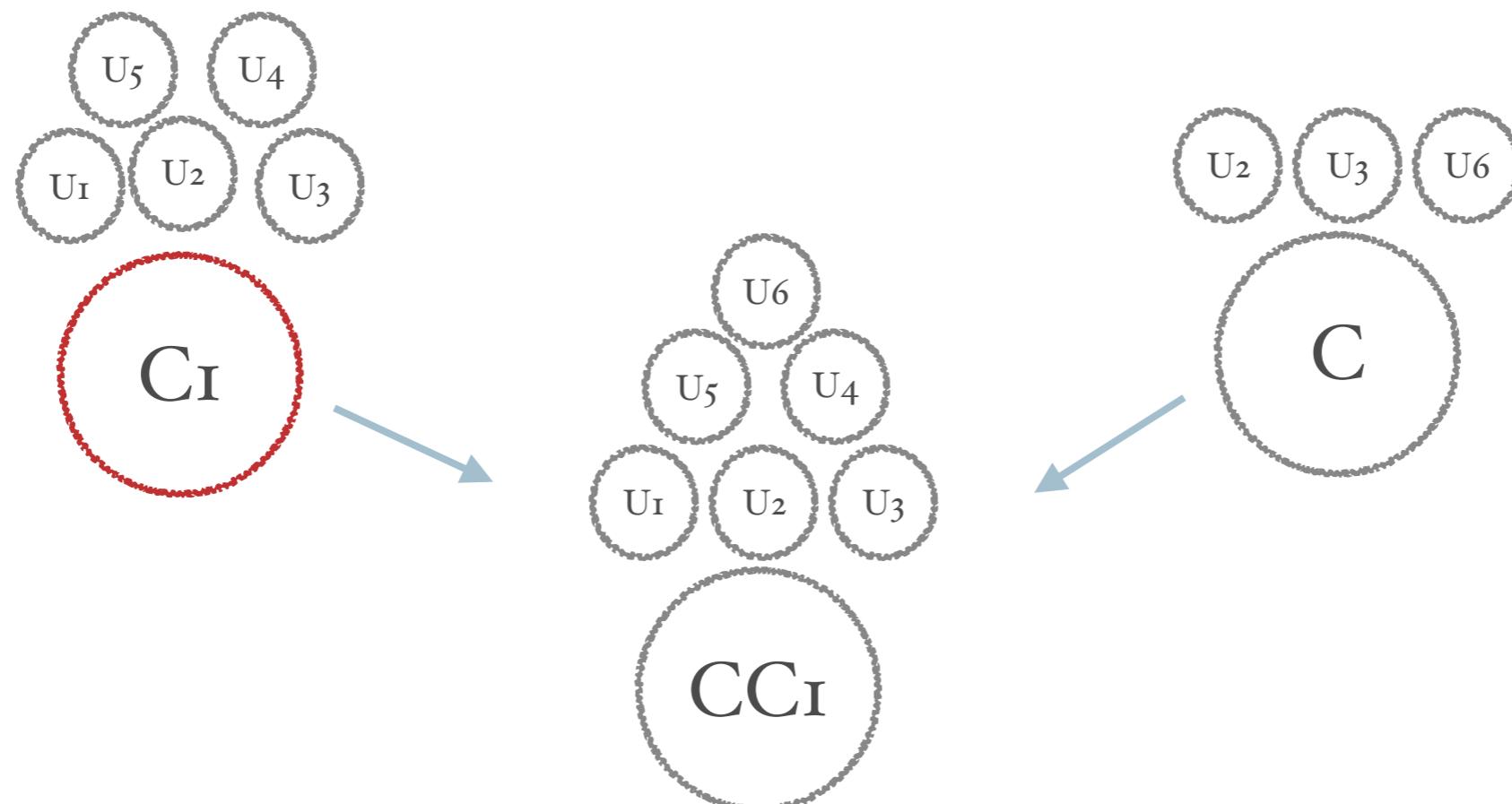
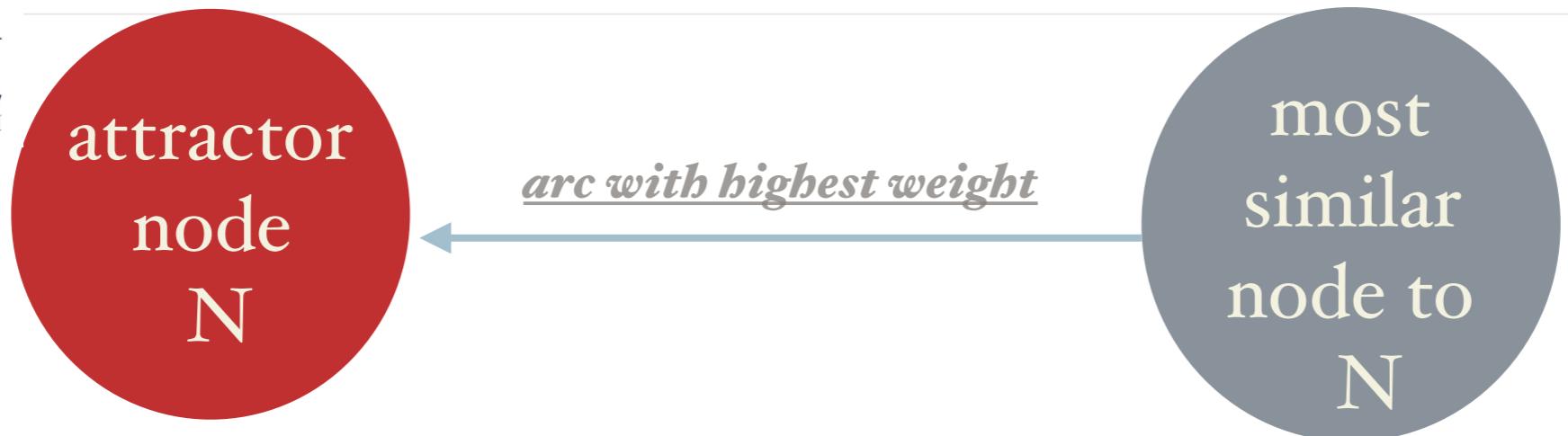
and
nodes

using semantic similarity between cores

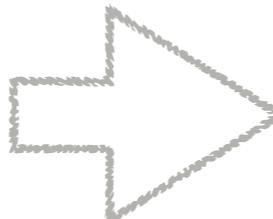
$$((1+\text{mean})^3)/\sqrt{\# \text{ of users}}$$

=

Attractor node

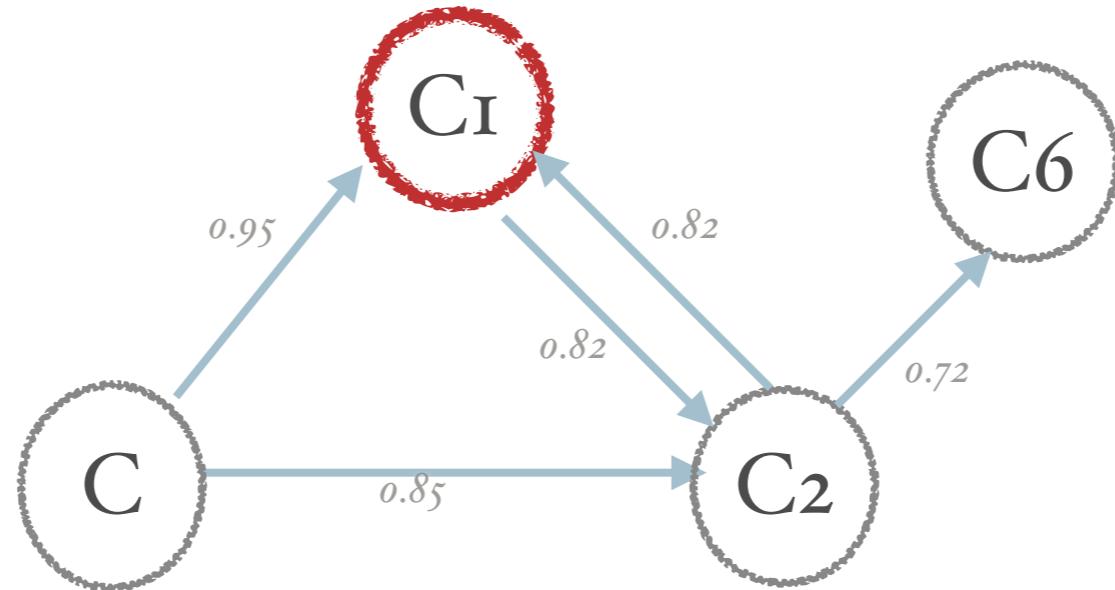


if $|CC_I$ ' s profile| > t

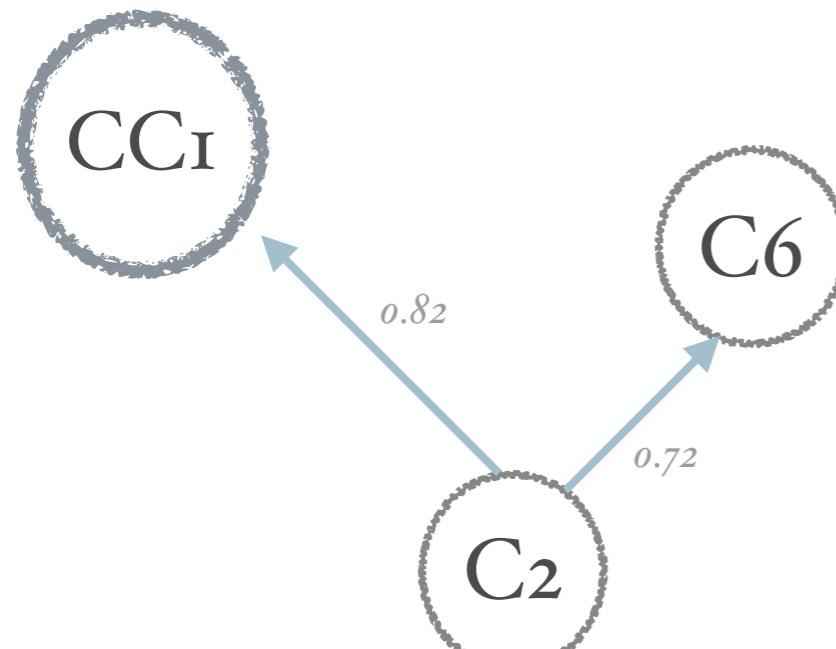


fusion confirmed

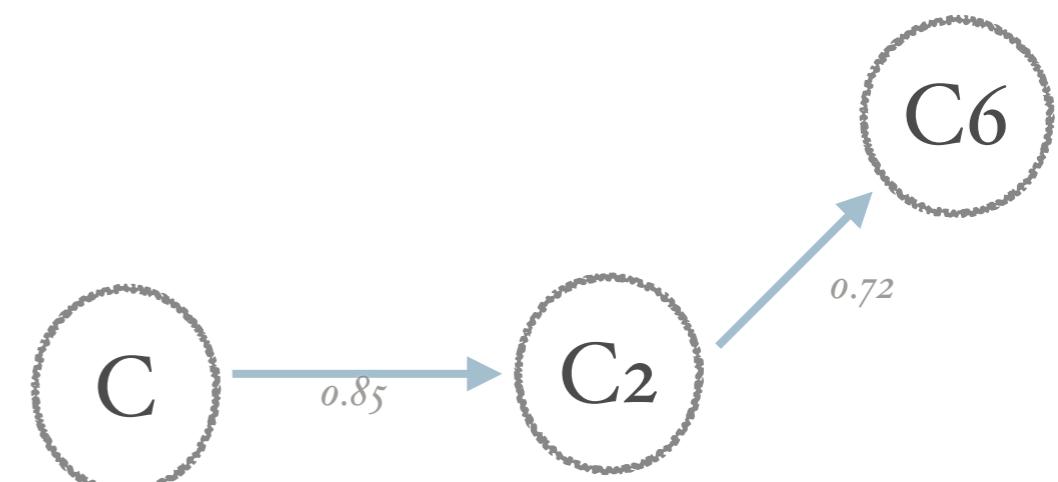
Updating proximity graph



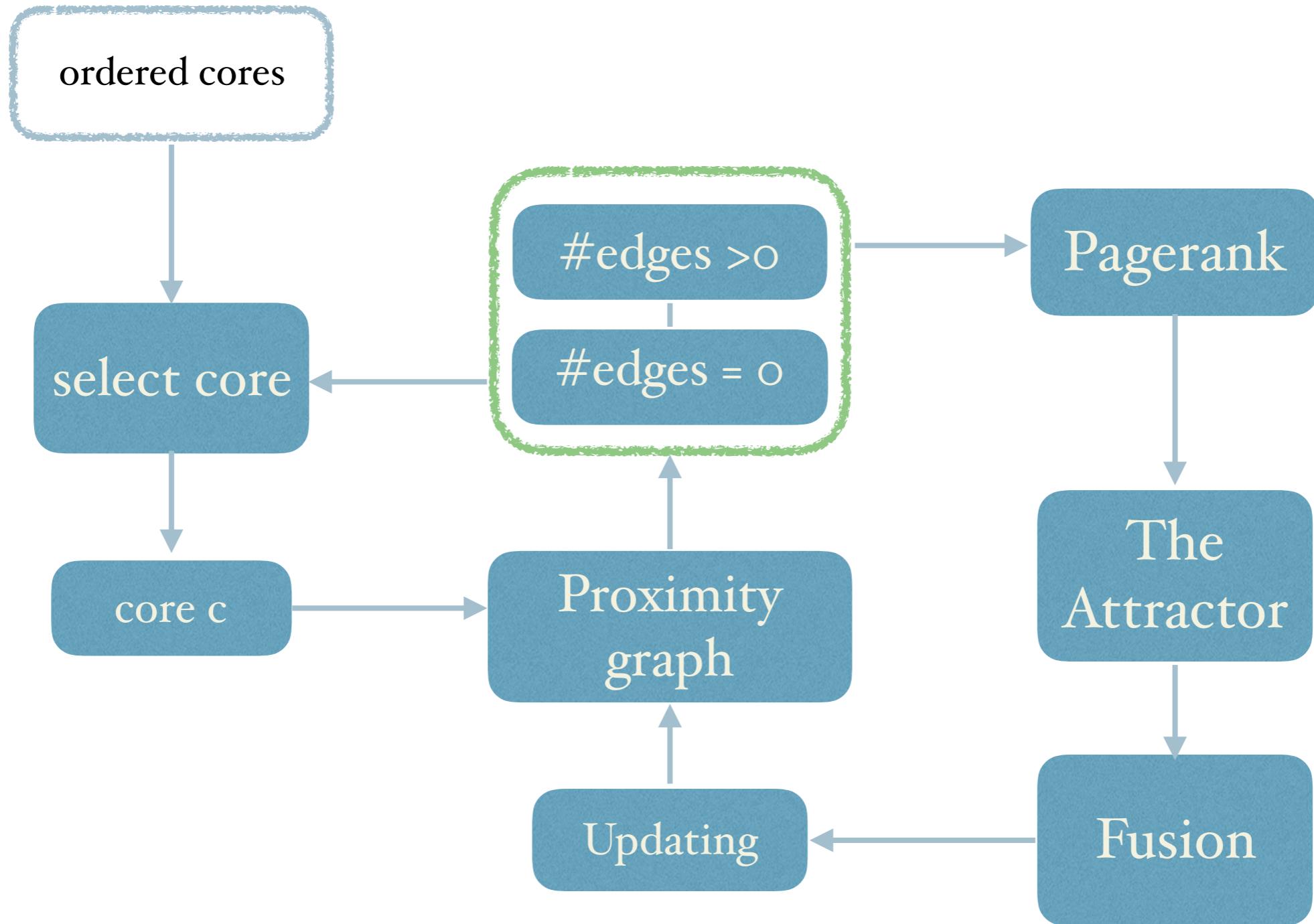
Approved 



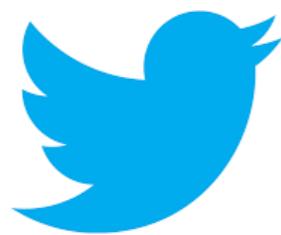
Not approved 



Modello di un'iterazione



Dataset



3500 users



36500 cliques

kaggle



friends +
followers > 200

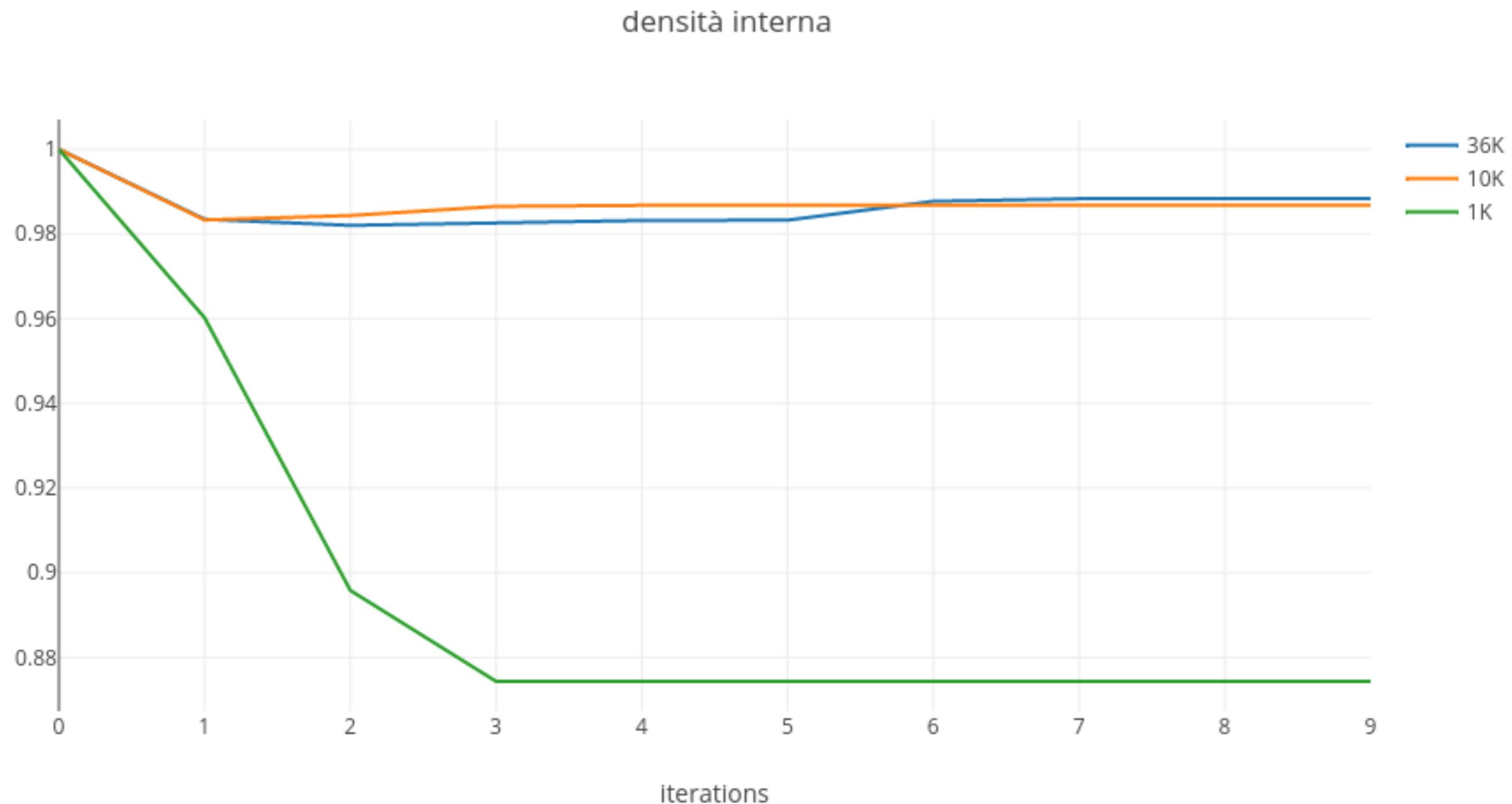
80 interest per
user



dense graph

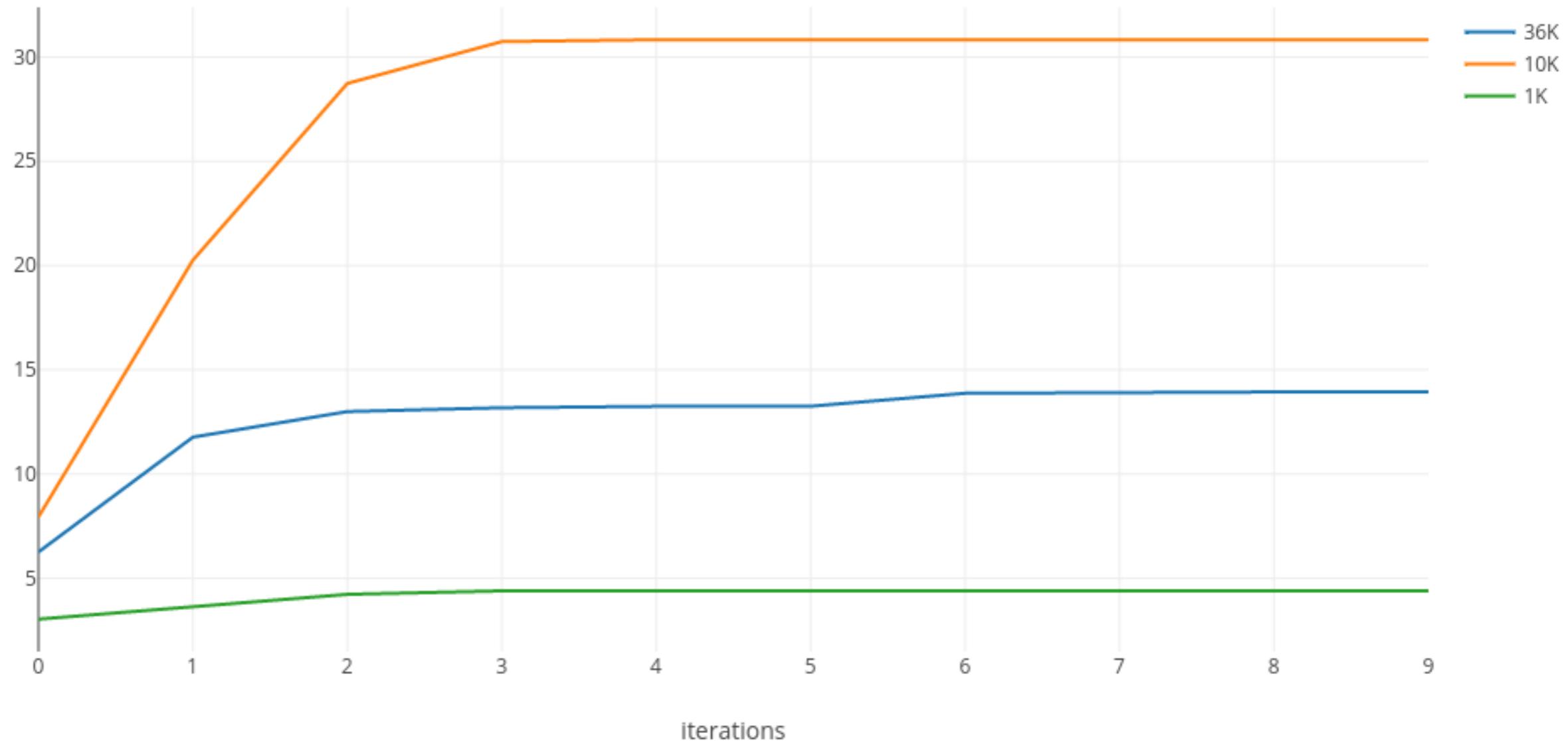
98 clique per
user

Effectiveness - internal density



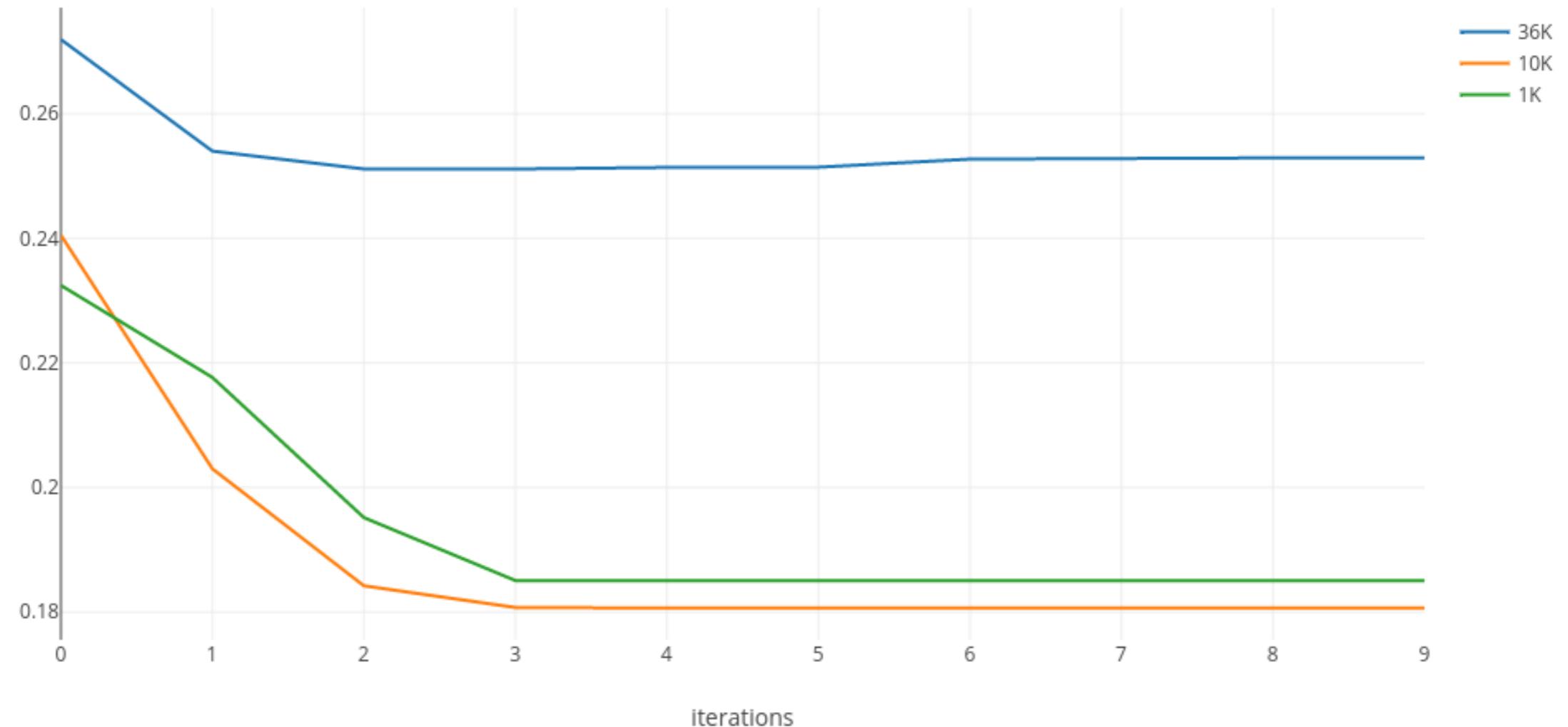
Effectiveness - mentions average in a community

numero mentions reciproche



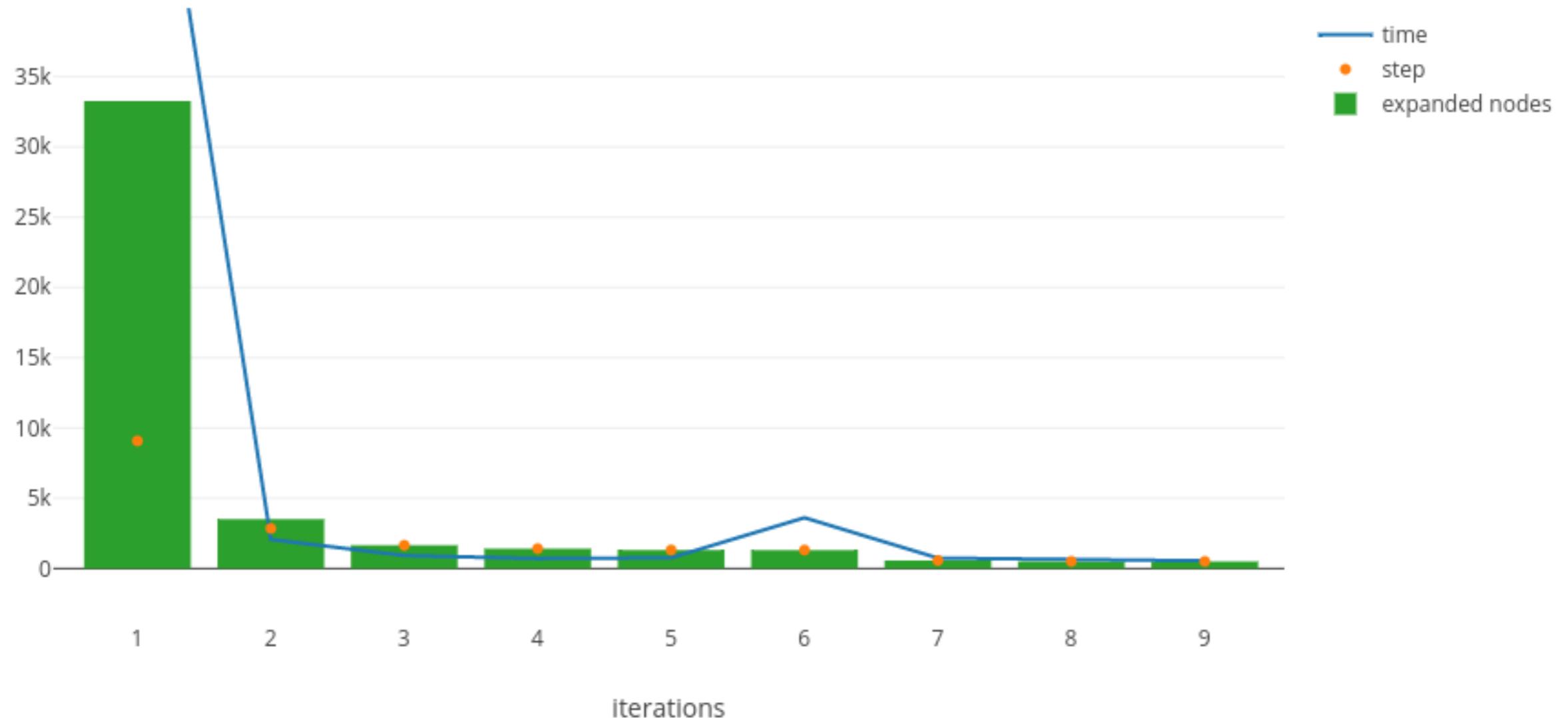
Effectiveness - profile overlapping with its community profile

personal profiles overlap cluster 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 parameter values*
- *using a different topological approach (different from clique) [in progress]*
- *test on different domains [in progress]*
- *model patterns on obtained outputs*
- *community prediction*

Contacts

- giorgiomaria1791@gmail.com
- <https://github.com/JoGreen>
- <https://www.linkedin.com/in/giorgio-maria-de-rango>

The end ..



