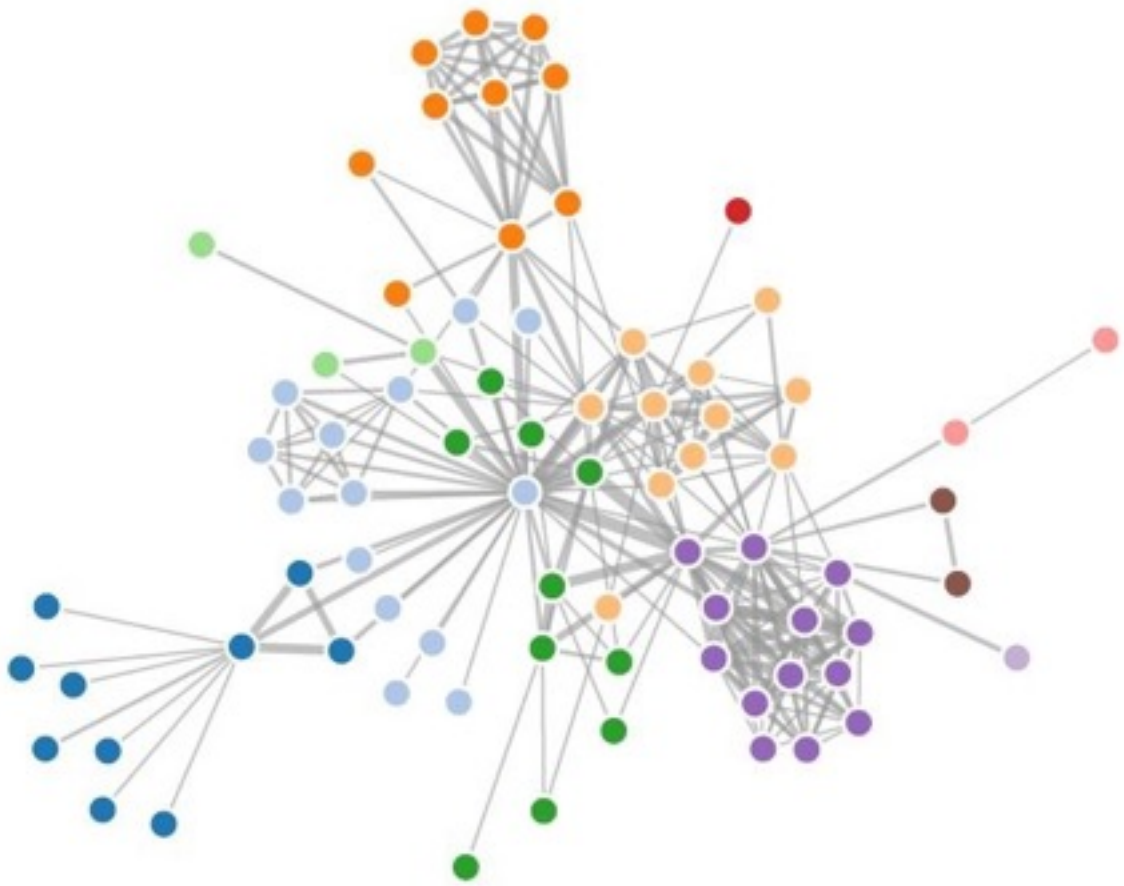
A Novel Approach to Evaluate Community Detection Algorithms on Ground Truth

Giulio Rossetti[1,2], Luca Pappalardo[1,2] and Salvatore Rinzivillo[2]

[1] Computer Science Dept. University of Pisa, Italy
[2] KDD Lab. ISTI-CNR, Pisa Italy



CompleNet - 23 March 2016

Community Discovery

High level meta-definition:

Given a graph provide a decomposition such that the nodes within each identified substructure are tightly connected within each other than with nodes belonging to a different one.

Problem(s):

A. Community Discovery Problem is *"ill posed"*:
there not exists a formal and shared definition of what a Community is;

B. How to evaluate a graph partition?

Partition Quality Evaluation

Two families of approaches to assess partition quality:

A. **Internal evaluation**

- Partition quality function
  (i.e., modularity, conductance, density…)
- Community characterization
  (i.e., size distribution, overlap distribution…)
- Execution time and Complexity

B. **External Evaluation**

- Ground truth testing

Ground Truth Testing

Idea

Given a graph G, a ground truth partition P(G) and the set of identified communities C estimate the resemblance the latter has with P(G).

General Criticism(s)

- Different approaches generates communities following different criteria (*"ill posed"* problem)
- It is not necessarily true that the ground truth represent the only valid semantic\topologic partition for the analyzed graph.

Classical Approach:
NMI (Normalized Mutual Information)

$$NMI(X,Y) = \frac{H(X)+H(Y)-H(X,Y)}{\frac{H(X)+H(Y)}{2}} \in [0,1]$$

is a measure of *similarity* borrowed from information theory: *H(X)* is the entropy of the random variable *X* associated to an identified community, *H(Y)* is the entropy of the random variable *Y* associated to a ground truth community, and *H(X,Y)* is the joint entropy.

Advantages

- Extensively used in literature

Drawbacks

- Computational complexity ~ $O(|C|^2)$
  (where C is the community set)
- Needs to be approximate in case of overlapping partitions

Our Proposal:
F1-Communities

Community Precision (P)
% of nodes in community ***x*** belonging to the ground truth community ***y***

$$P = \frac{|x \cap y|}{x}, \; P \in [0,1]$$

Community Recall (R)
% of nodes of the ground truth community ***y*** covered by ***x***

$$R = \frac{|x \cap y|}{y}, \; R \in [0,1]$$

F1-Community (F1)
harmonic mean of Precision and Recall

$$F1 = 2\frac{PR}{P+R}, \; F1 \in [0,1]$$

Advantages

- Computational complexity O(|C|):
  matching is performed by maximizing Precision
- Easy to interpret
- It is not affected by overlap

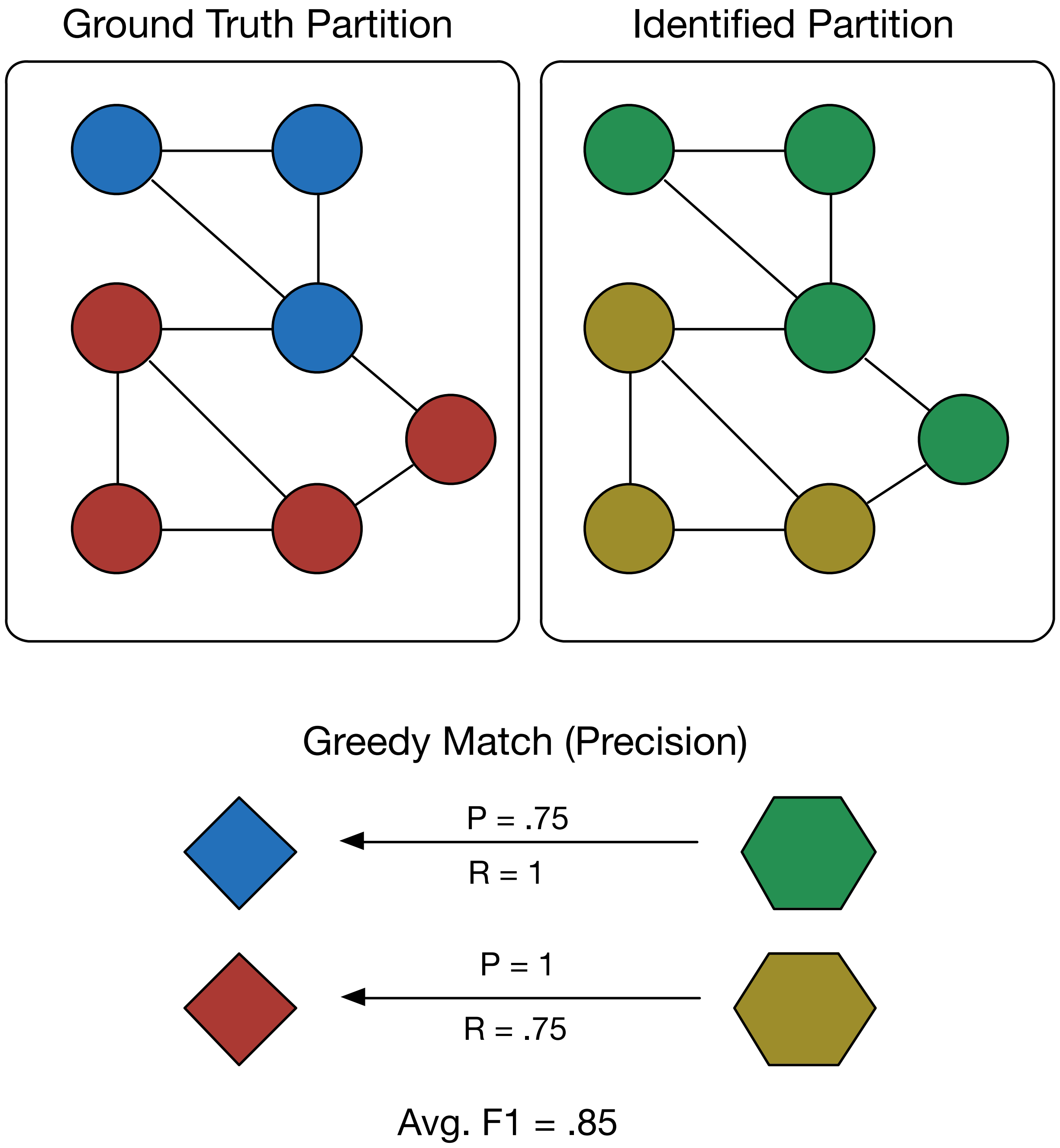
Example

1. Each node is labeled with its real communities

2. Each identified community is matched with the ground truth one whose label is shared by higher percentage of its nodes (higher Precision)

3. F1 is computed for each community

4. The average F1 score is given as quality score



Ground Truth Partition

Identified Partition

Greedy Match (Precision)

P = .75

R = 1

P = 1

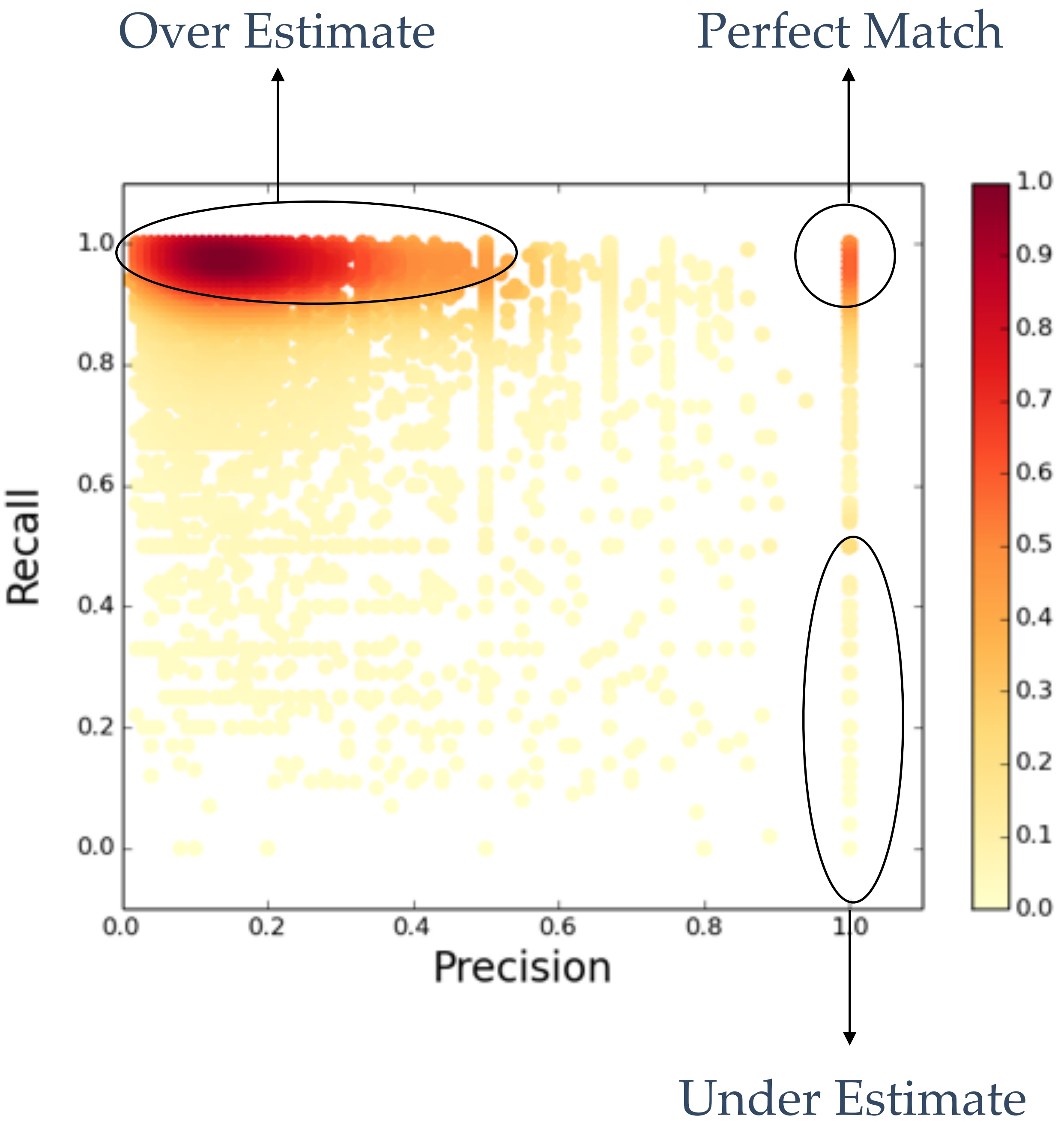
R = .75

Avg. F1 = .85

Visual Inspection

Precision and Recall allow to visualize the F1-matching quality with a density scatter plot:

- Max Precision & Max Recall imply a *perfect match*
- High Precision, Low Recall imply an *under estimate* of the real community
- High Recall, Low Precision imply an *over estimate* of the real community



[Extension]
Normalized F1-Communities

F1 is a conservative mean, however it can be not enough to accurately characterize the adherence of a graph partition to a ground truth.

We thus propose a normalized version:

$$NF1 = \frac{F1 * Coverage}{Redundancy}, \; NF1 \in [0, 1]$$

where given *P(G)*, C and $P_{id}(G)$

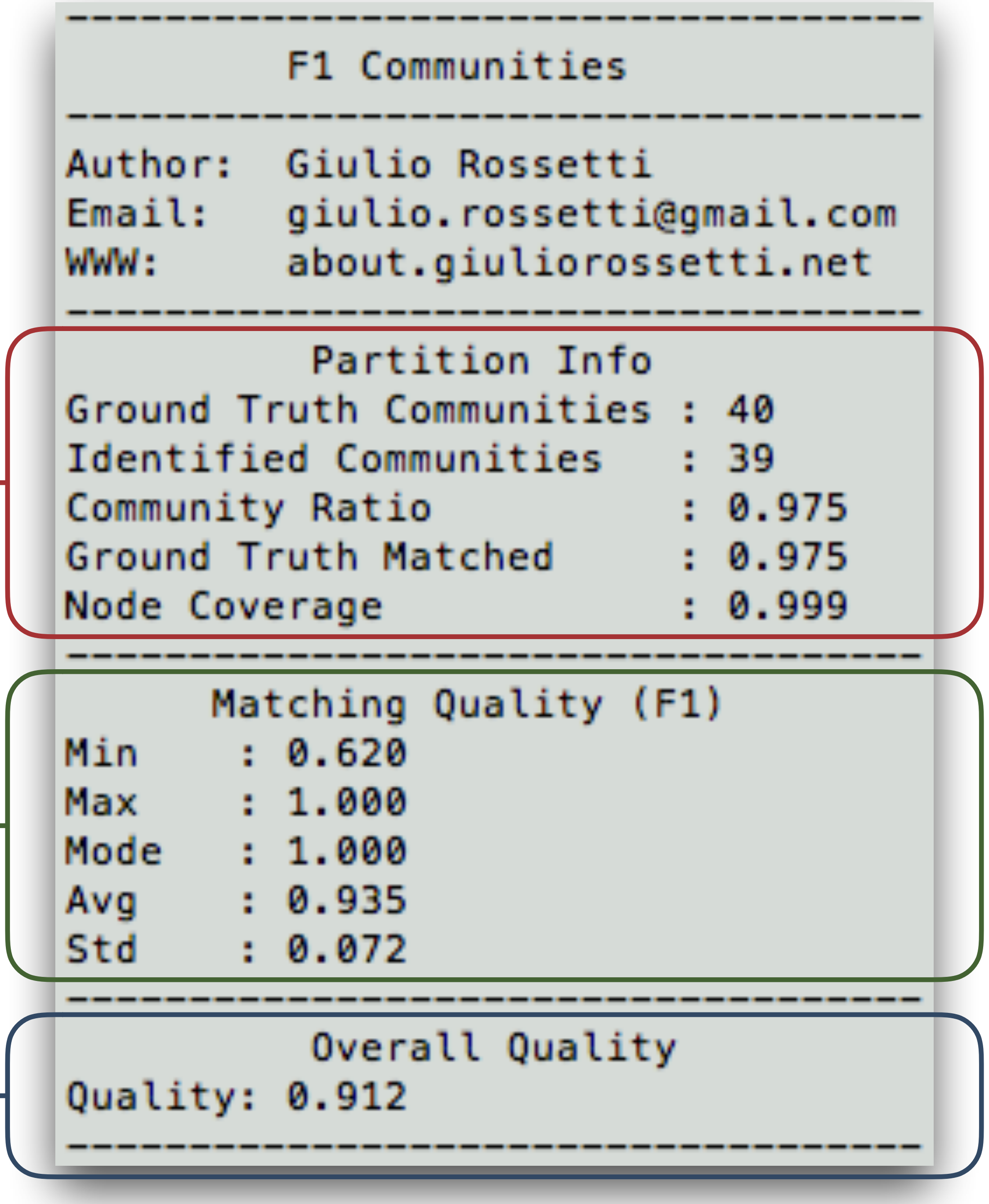
(i.e. the communities in *P(G)* matched by *C*)

$$Coverage = \frac{|P_{id}(G)|}{|P(G)|} \in [0, 1] \qquad Redundancy = \frac{|C|}{|P_{id}(G)|} \in [1, \infty]$$

Output Example

A Python implementation of F1-communities can be found at:

https://github.com/GiulioRossetti/f1-communities



```
-----------------------------------
          F1 Communities
-----------------------------------
Author:   Giulio Rossetti
Email:    giulio.rossetti@gmail.com
WWW:      about.giuliorossetti.net
-----------------------------------
           Partition Info
Ground Truth Communities : 40
Identified Communities   : 39
Community Ratio          : 0.975
Ground Truth Matched     : 0.975
Node Coverage            : 0.999
-----------------------------------
       Matching Quality (F1)
Min    : 0.620
Max    : 1.000
Mode   : 1.000
Avg    : 0.935
Std    : 0.072
-----------------------------------
          Overall Quality
Quality: 0.912
-----------------------------------
```

General information on community matches

F1 detailed statistics

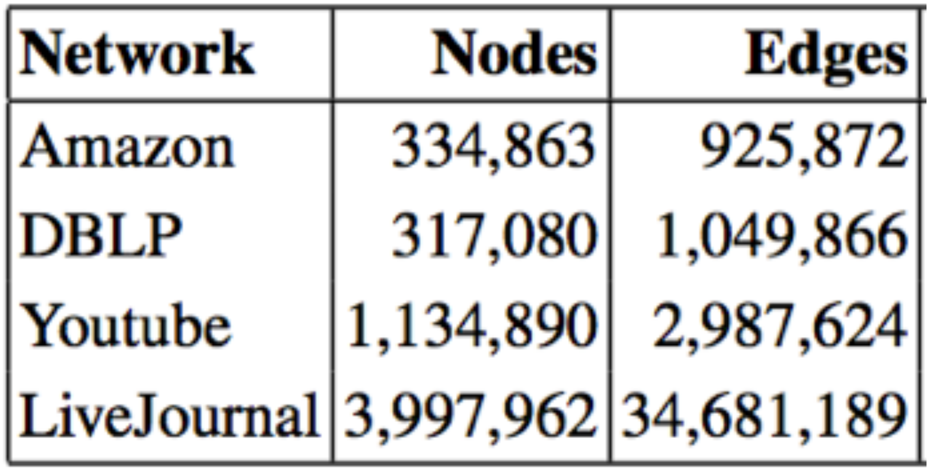
NF1 quality index

Experiments

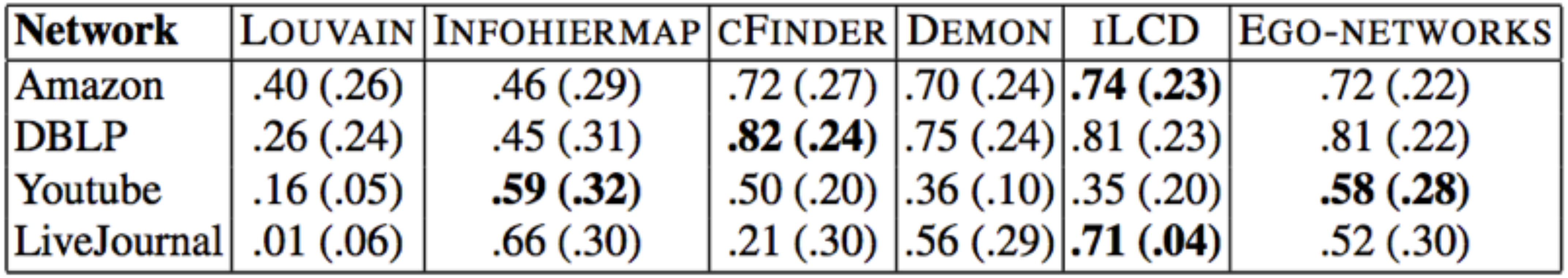
Six CD approaches tested…

- cFinder
- Louvain
- Infohiermap
- DEMON
- iLCD
- Ego-Networks

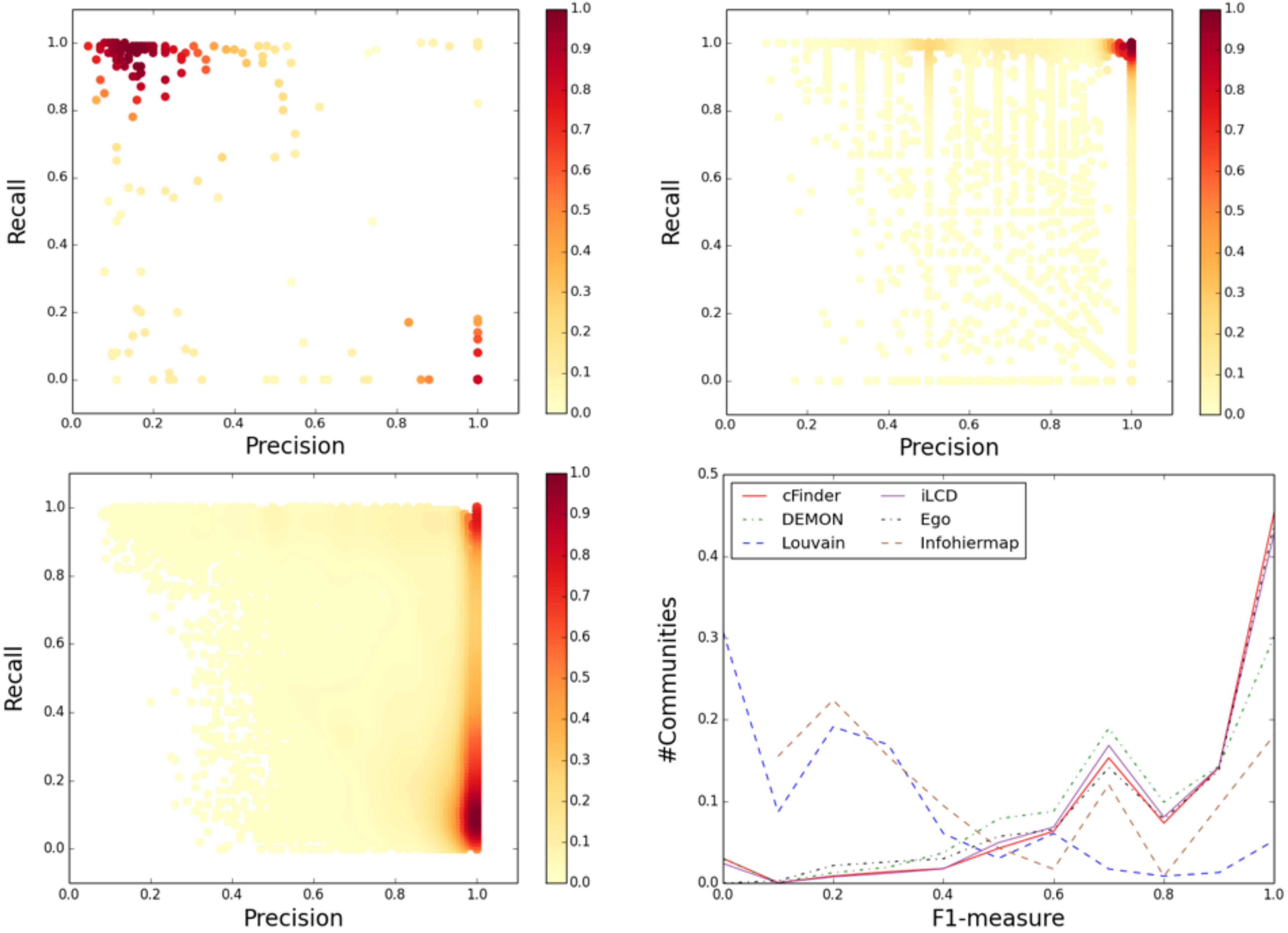
… on four real world graph with annotated ground truth communities.



Network	Nodes	Edges
Amazon	334,863	925,872
DBLP	317,080	1,049,866
Youtube	1,134,890	2,987,624
LiveJournal	3,997,962	34,681,189



Network	LOUVAIN	INFOHIERMAP	CFINDER	DEMON	ILCD	EGO-NETWORKS
Amazon	.40 (.26)	.46 (.29)	.72 (.27)	.70 (.24)	**.74 (.23)**	.72 (.22)
DBLP	.26 (.24)	.45 (.31)	**.82 (.24)**	.75 (.24)	.81 (.23)	.81 (.22)
Youtube	.16 (.05)	**.59 (.32)**	.50 (.20)	.36 (.10)	.35 (.20)	**.58 (.28)**
LiveJournal	.01 (.06)	.66 (.30)	.21 (.30)	.56 (.29)	**.71 (.04)**	.52 (.30)


Recall
Precision
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
Precision
Recall
Precision
#Communities
F1-measure
0.5
cFinder
iLCD
DEMON
Ego
Louvain
Infohiermap

Conclusions

NF1-communities is designed to:

- Evaluate the *degree of resemblance* among two graph partitions
- Allow for a *visual inspection* of the performances achieved by a community discovery algorithm
- *Reduce* the computation time of previous approaches (i.e. NMI)
- *Handle* overlap partitions (via redundancy and coverage)

Thanks for your attention!

Questions?

"A Novel Approach to Evaluate Community Detection Algorithms on Ground Truth"

Giulio Rossetti, Luca Pappalardo, Salvatore Rinzivillo

CompleNet 2016

Email: giulio.rossetti@di.unipi.it

F1: https://github.com/GiulioRossetti/f1-communities