Evaluating the scalability of clustering algorithms on protein-protein interaction networks

Jose Picado

Fall 2014

Network Methods in Bioinformatics

Outline

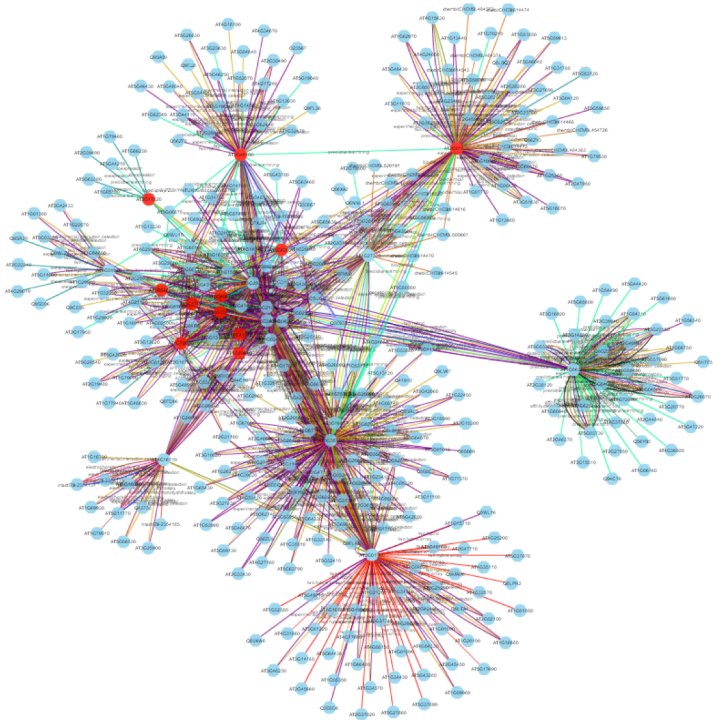
- Background
- Networks
- Algorithms
- Results
- Conclusions

Protein-protein interaction networks

- Structure of PPI networks reveals important properties
- Understand protein complexes
- Clustering:
  - Find sets of proteins that have more interactions among themselves
  - Clusters = protein complexes or functional modules

Protein-protein interaction networks

- PPI network = undirected graph
- Vertices = proteins
- Edges = interaction between proteins



Networks analyzed

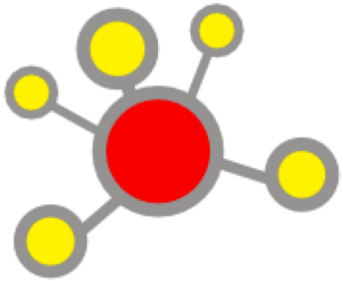
- BioGRID Database

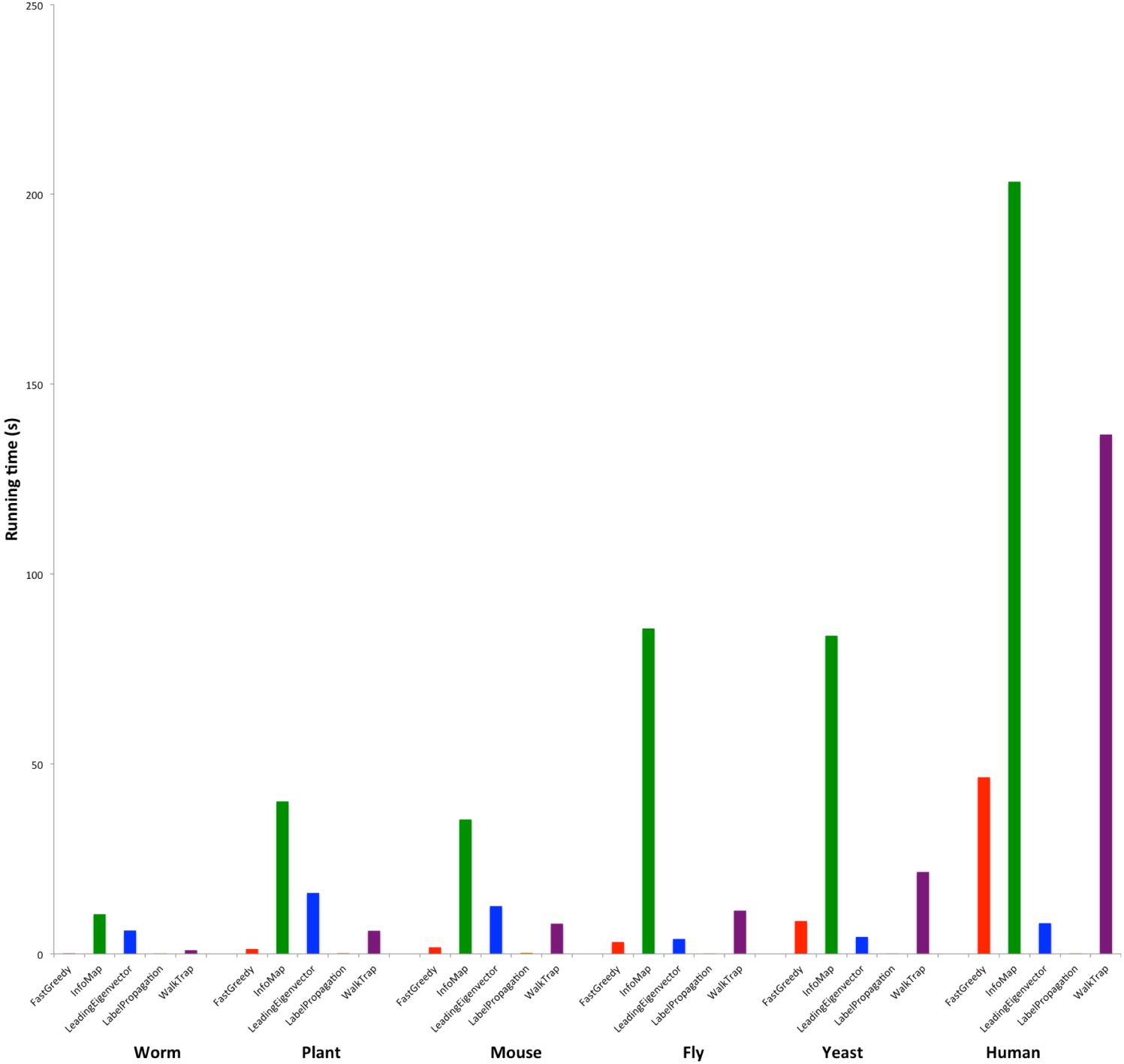
Organism	# of vertices	# of edges
Worm	3288	6353
Plant	7200	21536
Mouse	8317	22820
Fly	8077	37611
Yeast	6341	138856
Human	187747	243671

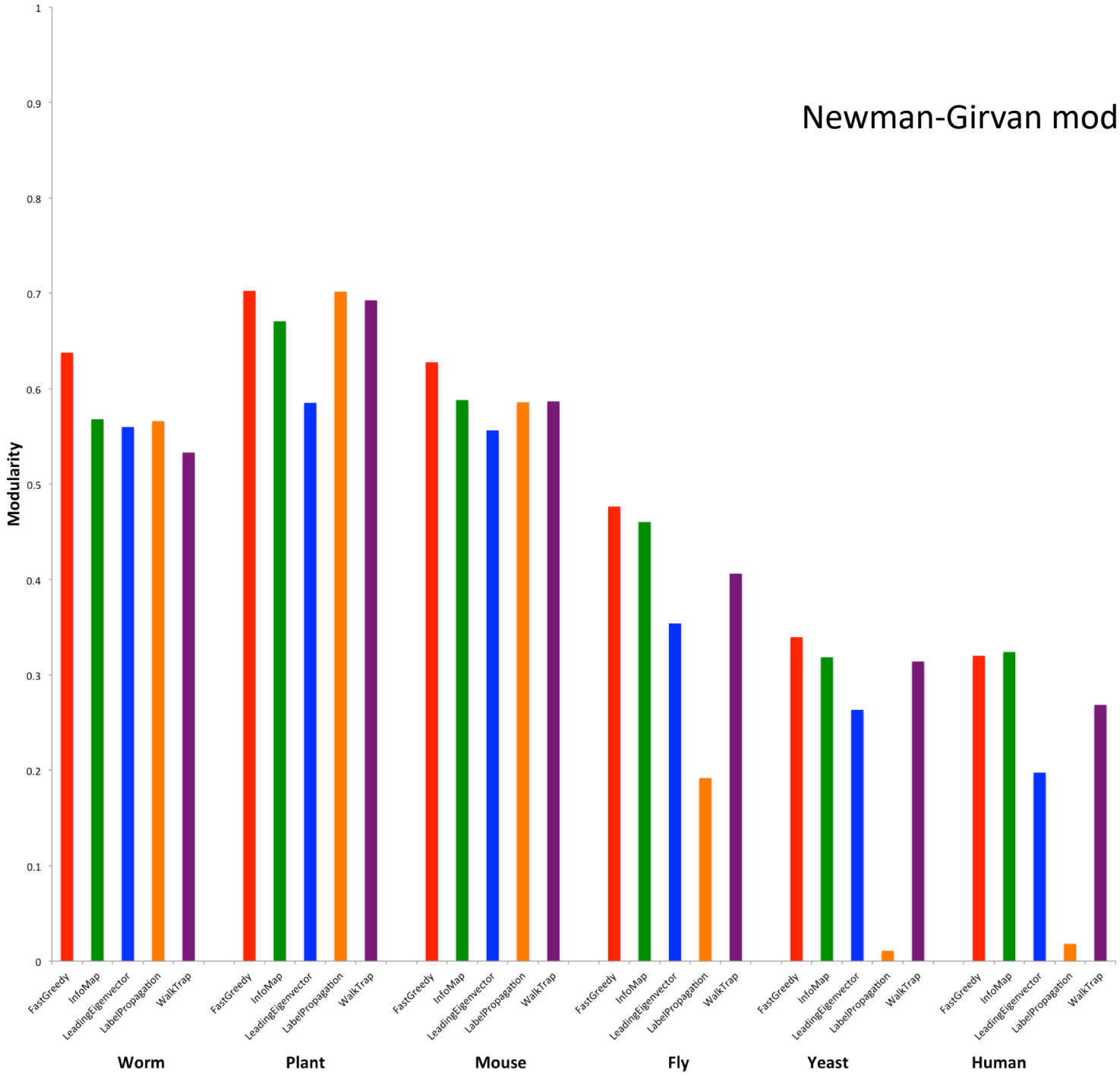
Clustering algorithms

- Walktrap
- Leading Eigenvectors
- Fast Greedy
- Label Propagation
- InfoMAP

- All implemented in igraph
- Code developed in python




Running time (s)
250
200
150
100
50
0
FastGreedy
InfoMap
LeadingEigenvector
LabelPropagation
WalkTrap
Worm
Plant
Mouse
Fly
Yeast
Human


Newman-Girvan modularity
Modularity
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
FastGreedy
InfoMap
LeadingEigenvector
LabelPropagation
WalkTrap
Worm
Plant
Mouse
Fly
Yeast
Human

Modularity scores

- Fast-Greedy: best trade-off between efficiency and effectiveness
- 100 random networks: shuffle associations in edge list while keeping degrees

Organism	Real network	Random networks
Worm	0.64	0.48
Plant	0.70	0.34
Mouse	0.63	0.36
Fly	0.48	0.25
Yeast	0.34	0.09
Human	0.32	0.13

Conclusions

- All algorithms are affected by network size
- Some algorithms are more scalable in terms of:
  - Efficiency (running time)
  - Effectiveness (modularity)
- Future work
  - Analyze algorithms in depth
  - Analyze scalability based on other properties

Thank you