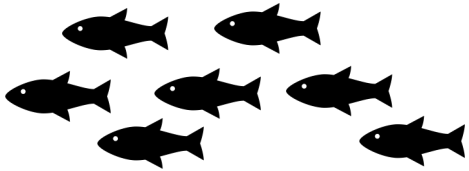
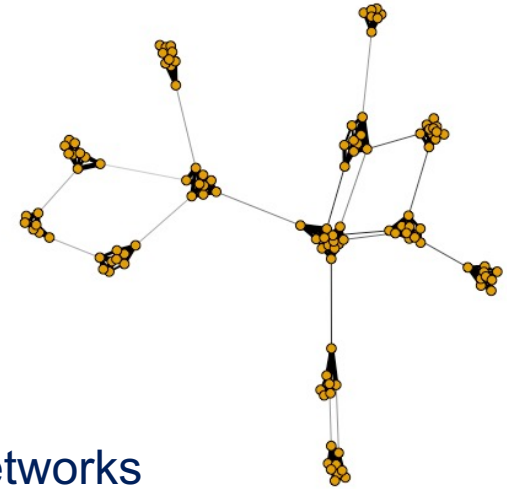
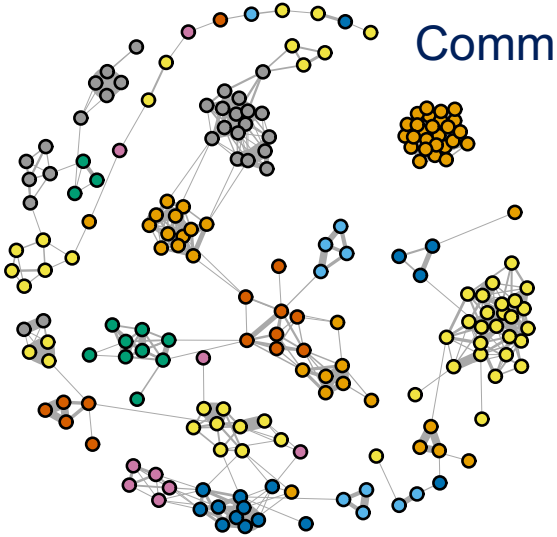
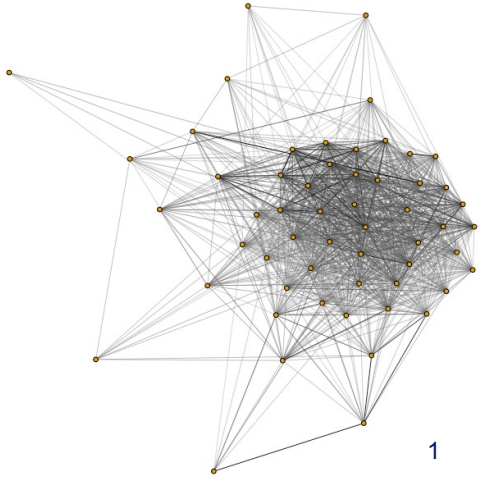




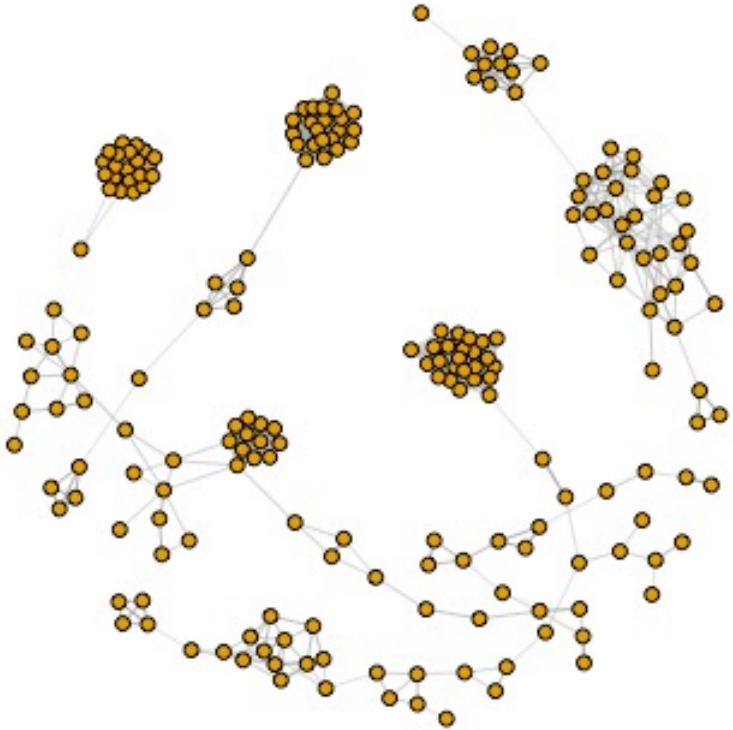
ExoLabel:
Community Detection for Big Biological Networks

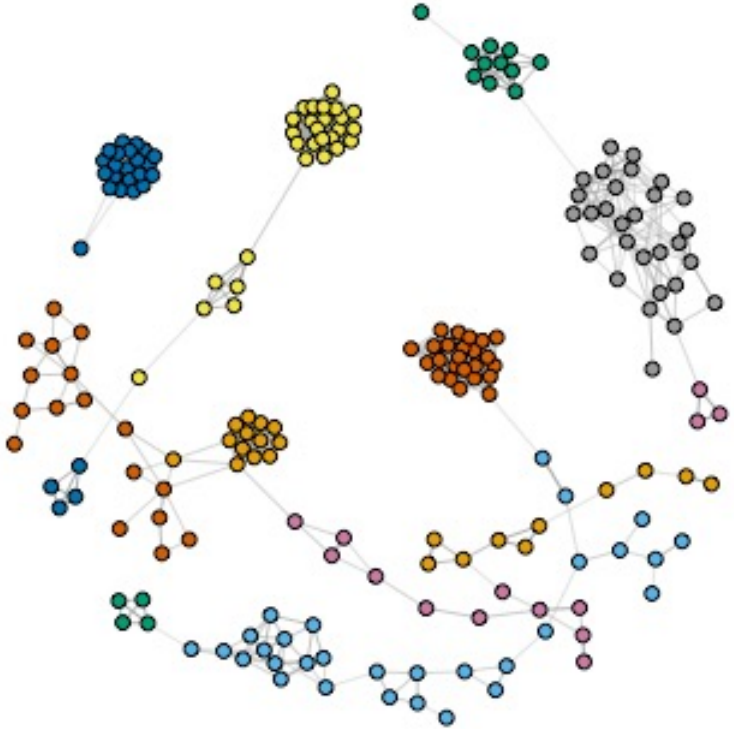
Aidan Lakshman, Erik S. Wright



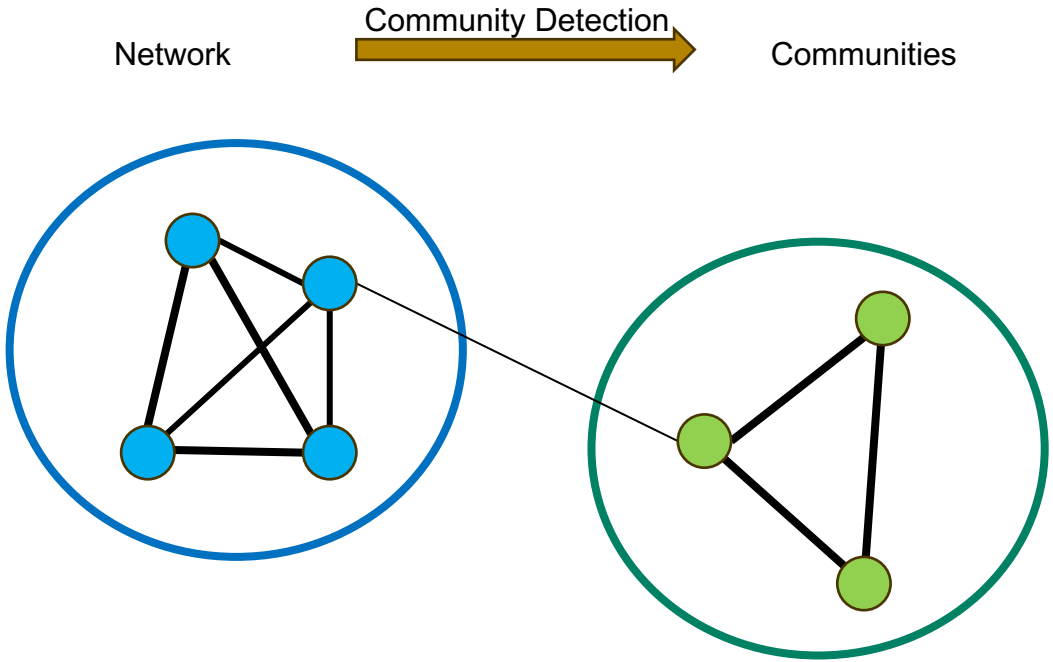


University of Pittsburgh

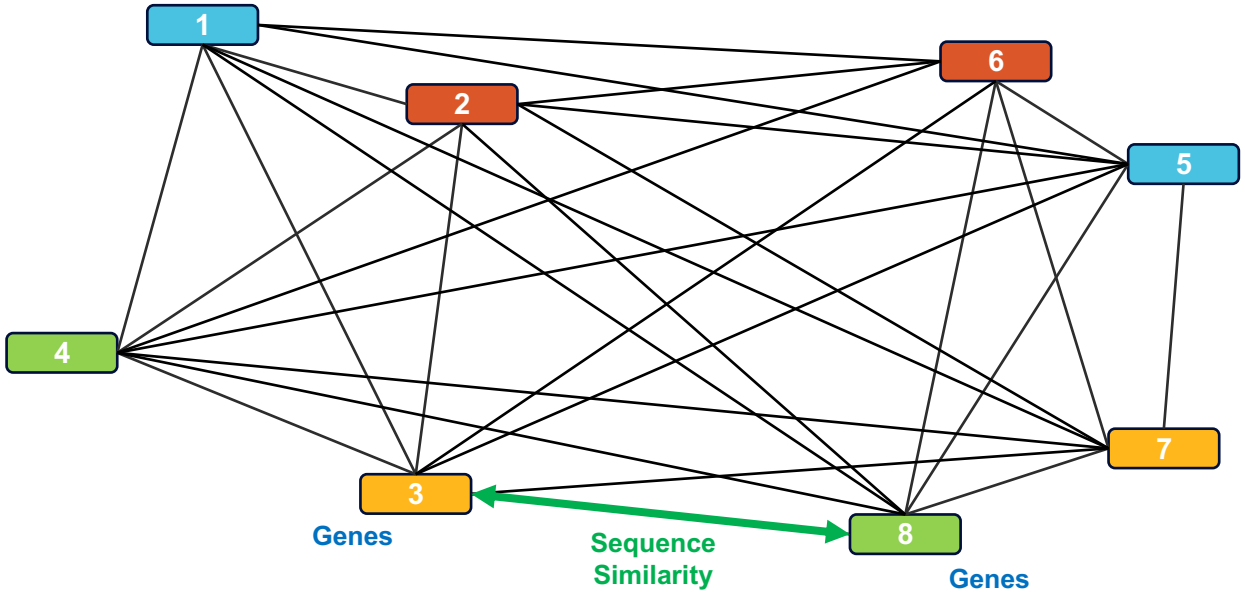




University of Pittsburgh


Network
Community Detection
Communities

Comparative genomics depends on detecting community detection in **Sequence Similarity Networks**

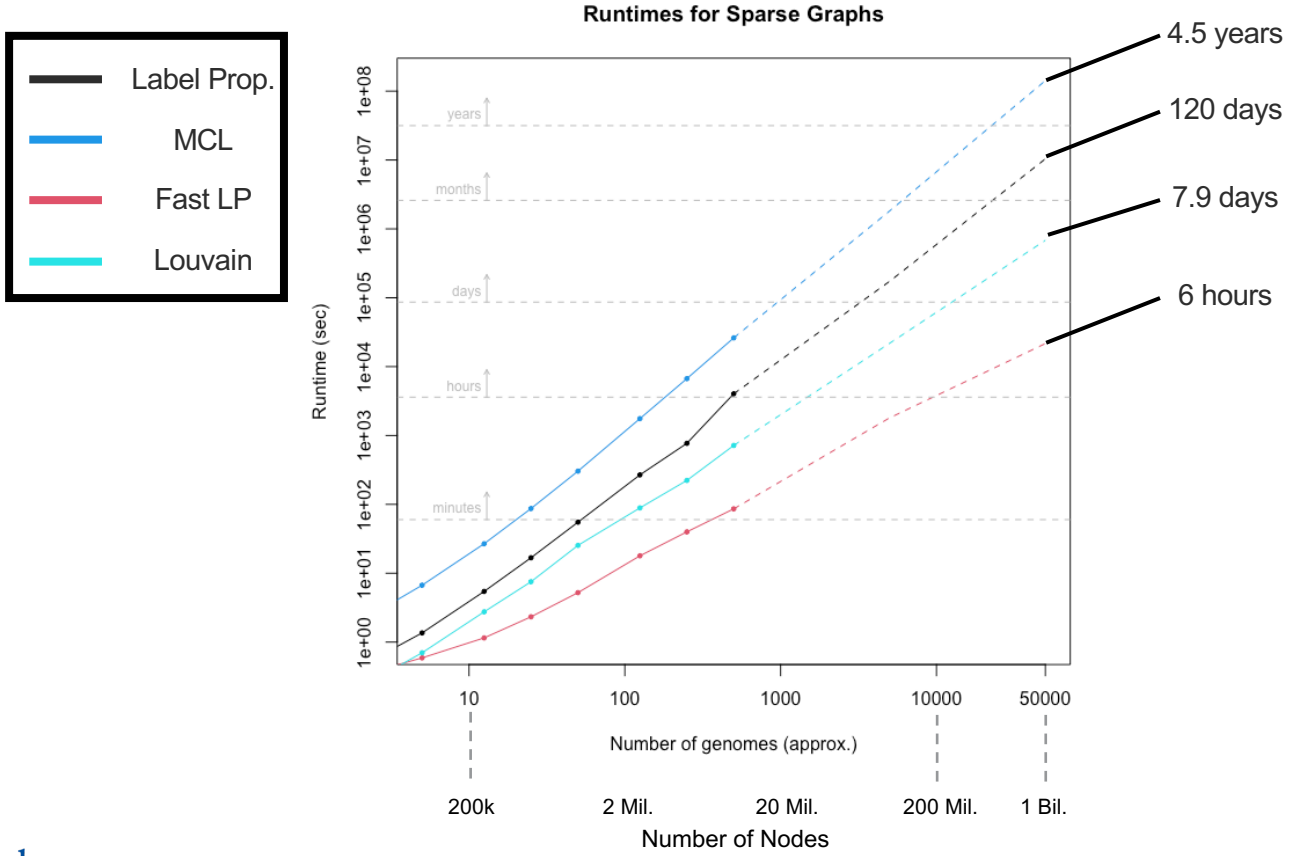


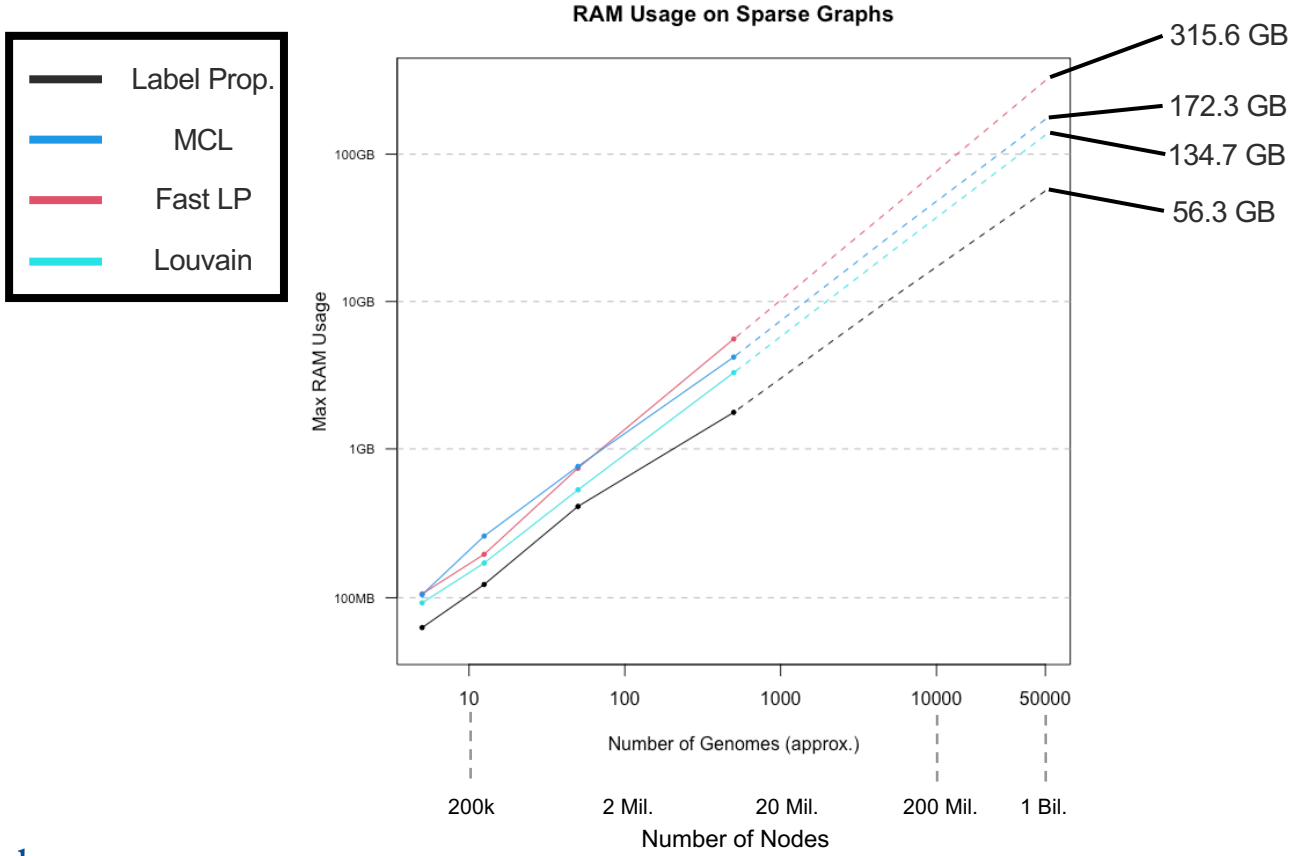
Thousands of genes

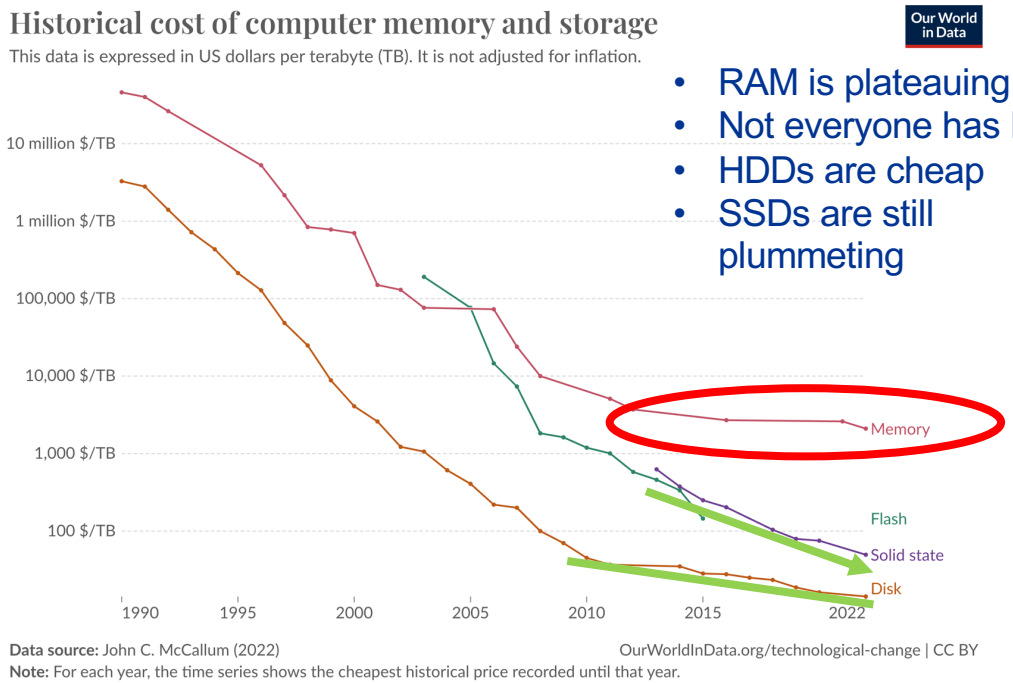
Each additional gene

1 edge per existing genome

Network size is quadratic in number of genomes

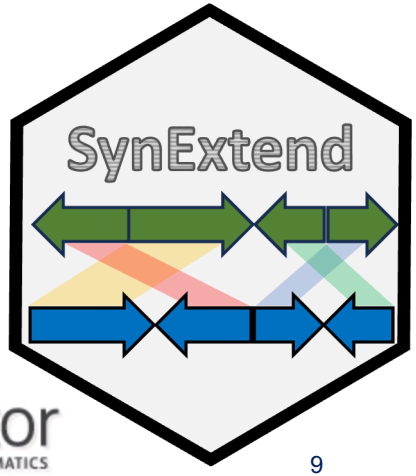

Label Prop.
MCL
Fast LP
Louvain
Runtimes for Sparse Graphs
Runtime (sec)
1e+00
1e+01
1e+02
1e+03
1e+04
1e+05
1e+06
1e+07
1e+08
years
months
days
hours
minutes
4.5 years
120 days
7.9 days
6 hours
10
100
1000
10000
50000
Number of genomes (approx.)
200k
2 Mil.
20 Mil.
200 Mil.
1 Bil.
Number of Nodes


Label Prop.
MCL
Fast LP
Louvain
RAM Usage on Sparse Graphs
315.6 GB
172.3 GB
134.7 GB
56.3 GB
100GB
10GB
1GB
100MB
Max RAM Usage
10
100
1000
10000
50000
Number of Genomes (approx.)
200k
2 Mil.
20 Mil.
200 Mil.
1 Bil.
Number of Nodes



- RAM is plateauing
- Not everyone has HPCs
- HDDs are cheap
- SSDs are still plummeting

ExoLabel



University of Pittsburgh

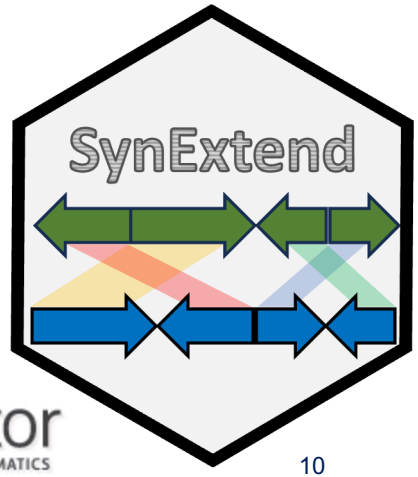
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Core algorithm:
Fast Label Propagation

Exo

Data stored out of memory

Label

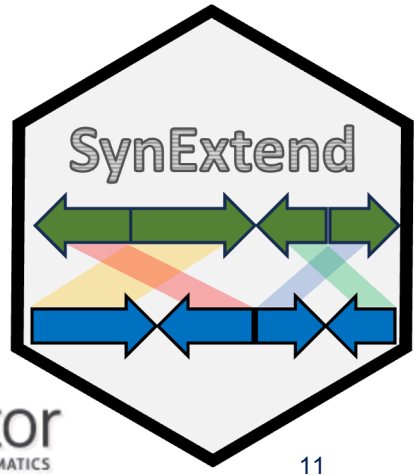


University of Pittsburgh

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

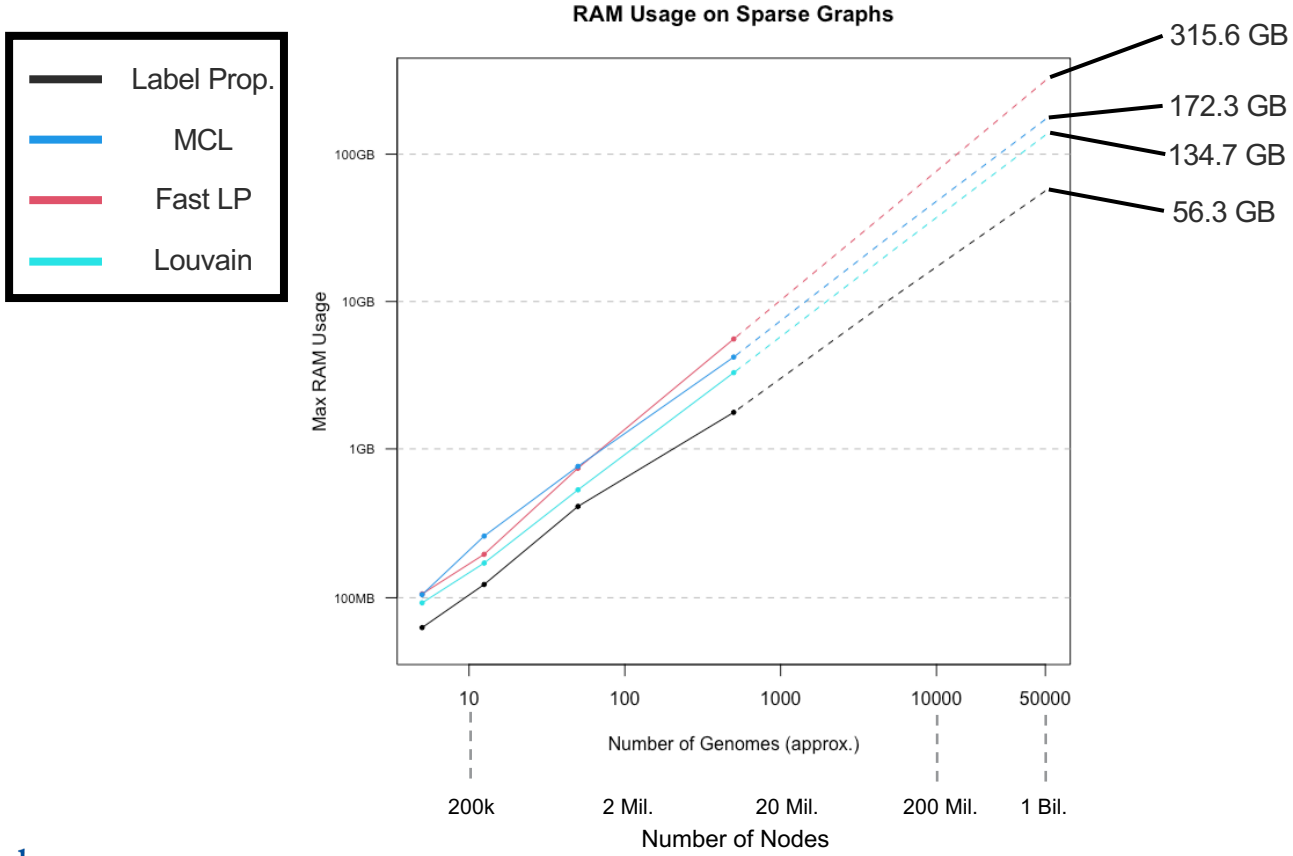
ExoLabel

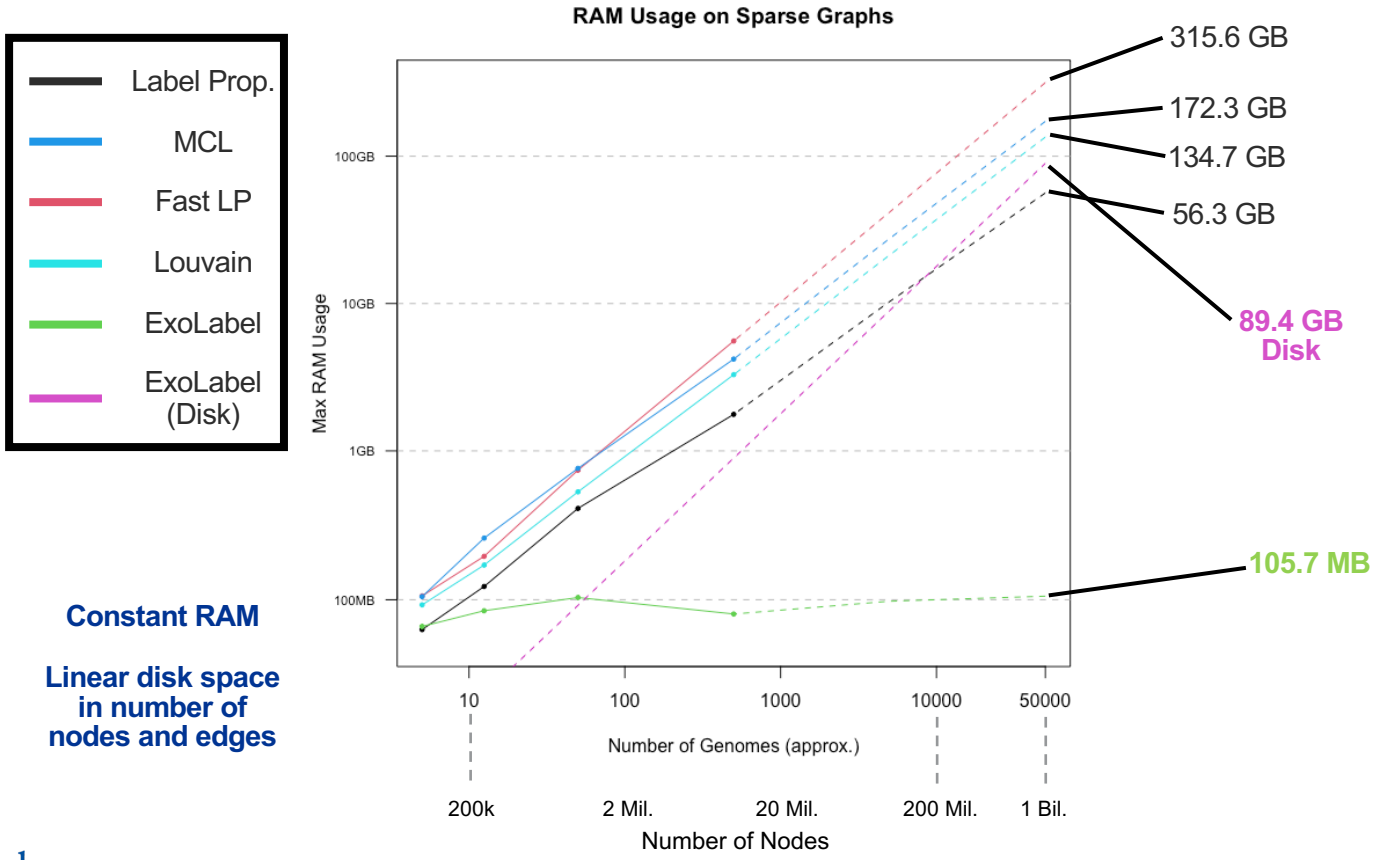
- Constant memory scaling
- Linear disk scaling
- Log-linear runtime
- Comparable accuracy to other standard methods



University of Pittsburgh

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

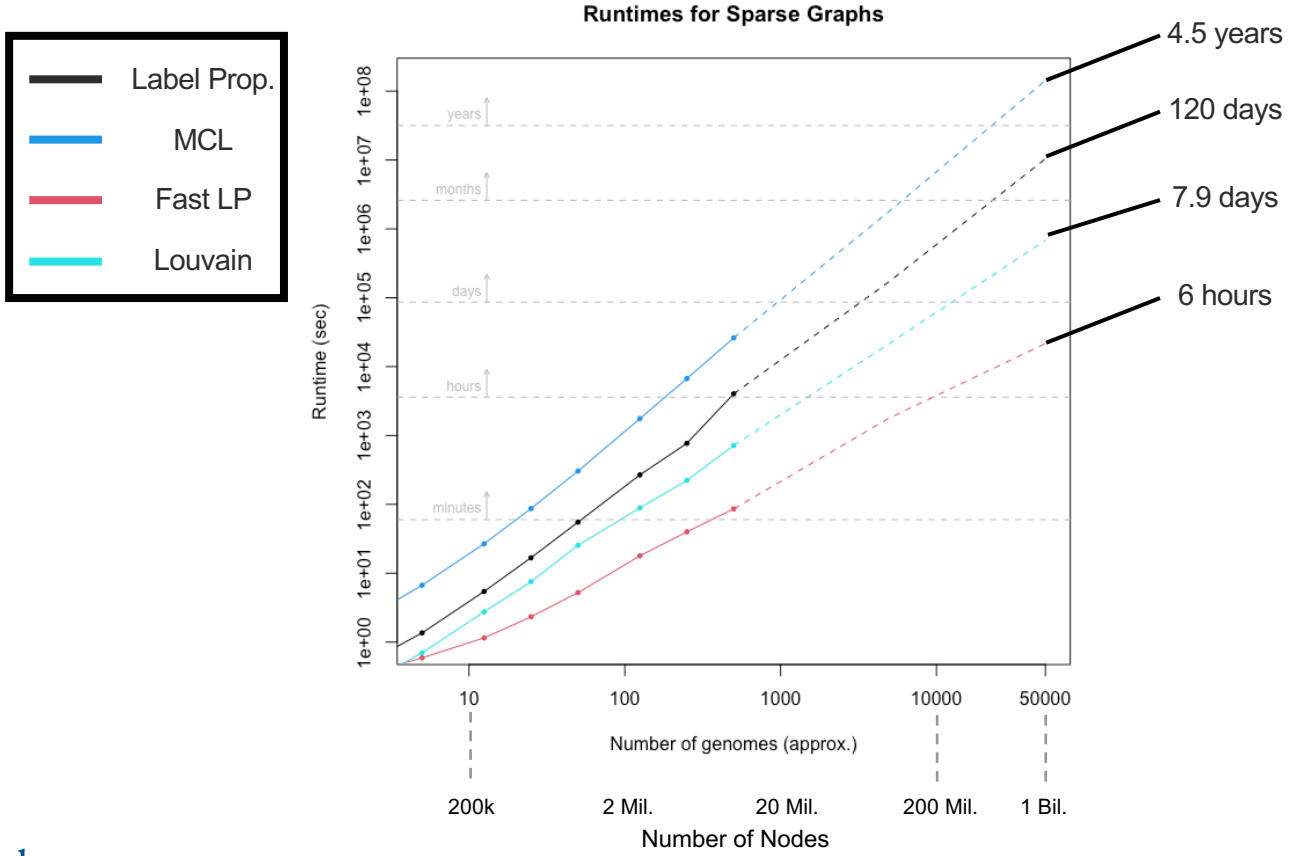

Label Prop.
MCL
Fast LP
Louvain
RAM Usage on Sparse Graphs
315.6 GB
172.3 GB
134.7 GB
56.3 GB
100GB
10GB
1GB
100MB
Max RAM Usage
10
100
1000
10000
50000
Number of Genomes (approx.)
200k
2 Mil.
20 Mil.
200 Mil.
1 Bil.
Number of Nodes

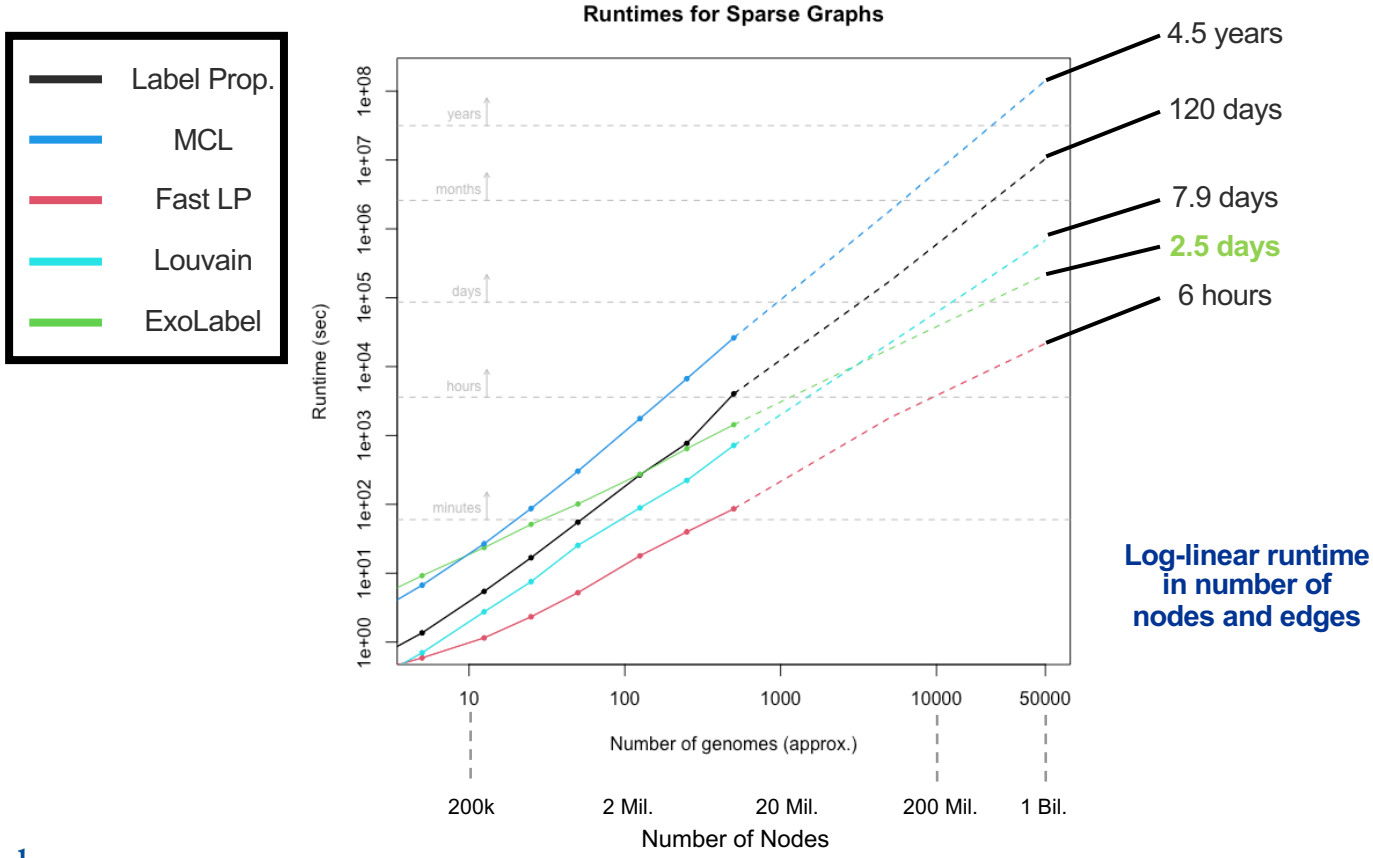


Constant RAM

Linear disk space in number of nodes and edges

University of Pittsburgh


Label Prop.
MCL
Fast LP
Louvain
Runtimes for Sparse Graphs
4.5 years
120 days
7.9 days
6 hours
Runtime (sec)
1e+00
1e+01
1e+02
1e+03
1e+04
1e+05
1e+06
1e+07
1e+08
years
months
days
hours
minutes
10
100
1000
10000
50000
Number of genomes (approx.)
200k
2 Mil.
20 Mil.
200 Mil.
1 Bil.
Number of Nodes



Log-linear runtime in number of nodes and edges

Acknowledgements

Erik Wright, PhD

- Nick Cooley, PhD
- Monica Tomaszewski, PhD
- Shu-Ting Cho
- Sam Blechman
- Nishant Panicker
- Maria Bond

www.WrightLabScience.com

AHL27.com
AHL27@pitt.edu

University of Pittsburgh

Research reported in this work was supported by the National Library of Medicine of the National Institutes of Health under award number **5T15LM007059-35**.