Network Analysis on Phylogenetic Data

Jeremy M. Brown [1], Guifang Zhou [1], Jeremy Ash [1], Wen Huang [2], Melissa Marchand [3], Kyle A. Gallivan [3], Jim C. Wilgenbusch [4]

[1]Department of Biological Sciences, Louisiana State University
[2]ICTEAM Institute, Université catholique de Louvain
[3]Department of Mathematics, Florida State University
[4]Minnesota Supercomputing Institute, University of Minnesota

October 26, 2015

Outline

Networks

What are networks?

Computer Networks

Figure: A small computer network

Social Networks

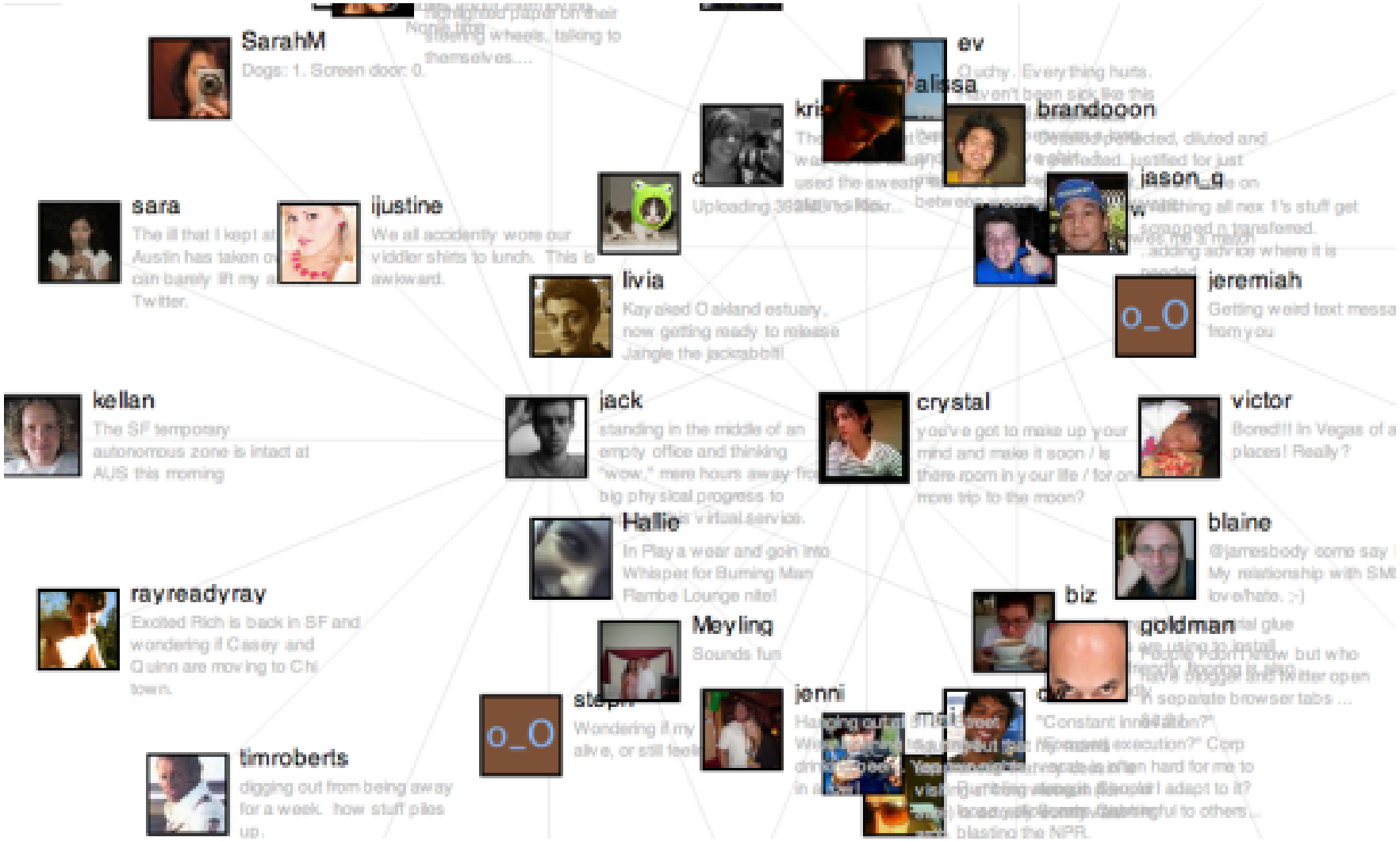


Figure: Twitter brower as a network of interconnections.

Collaboration Networks

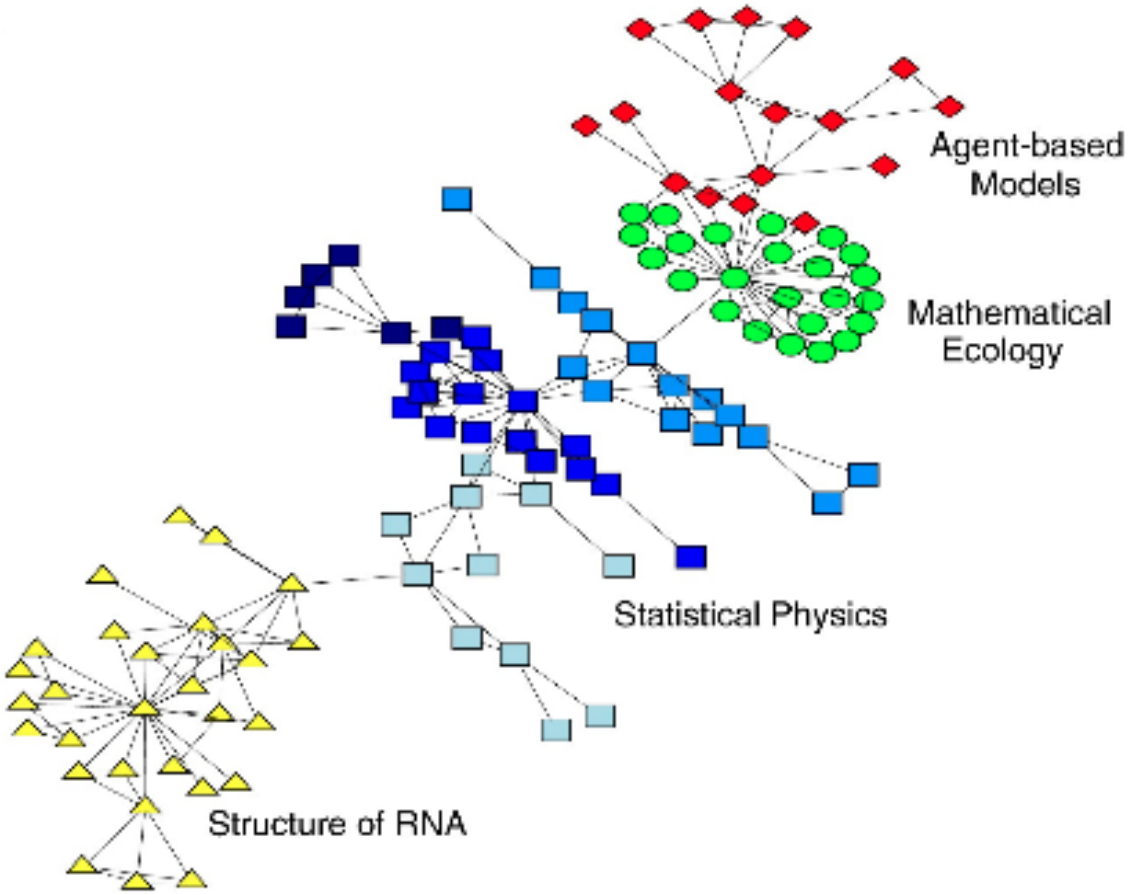


Figure: Collaboration network between scientists working in Santa Fe Institute.

Biology Networks

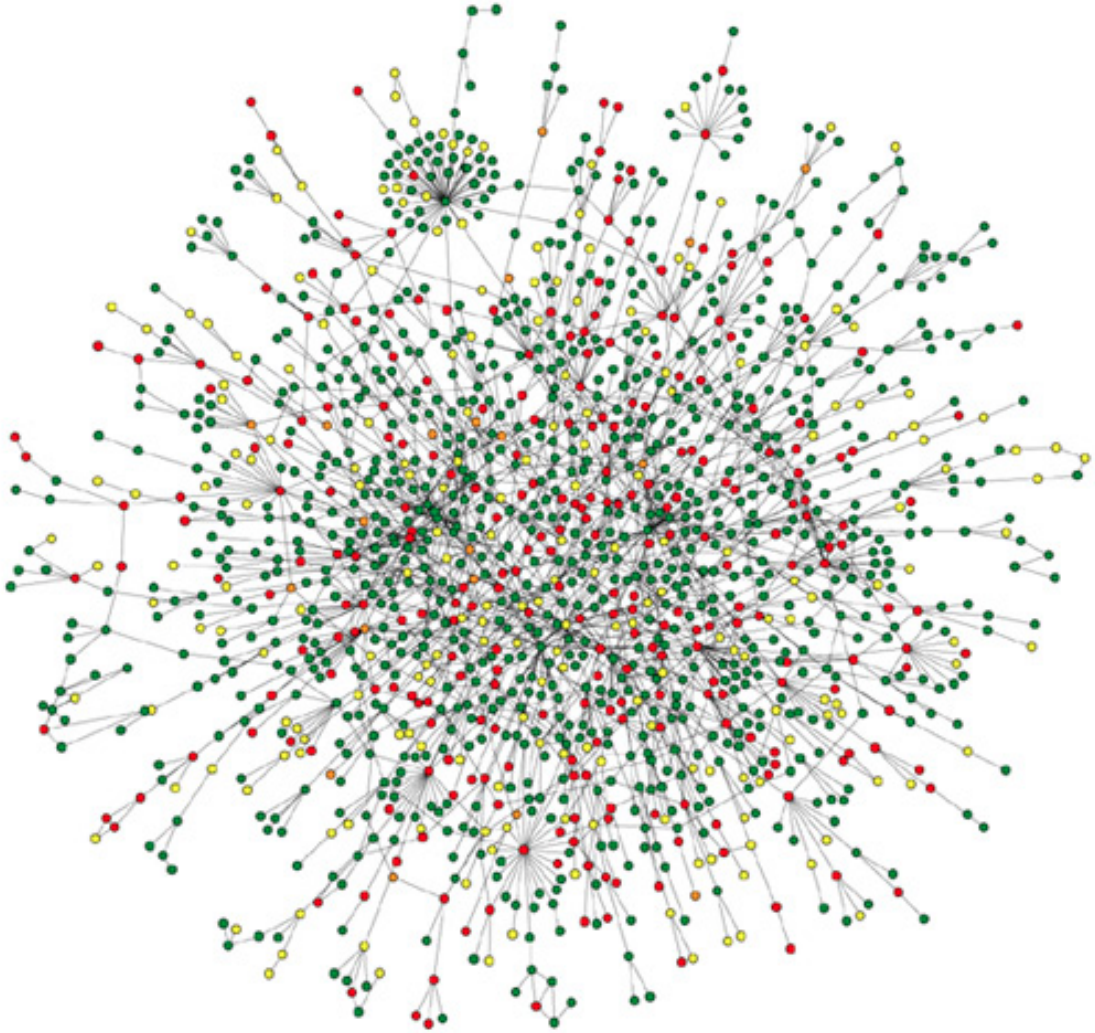


Figure: Yeast protein interaction network

Networks of Trees?

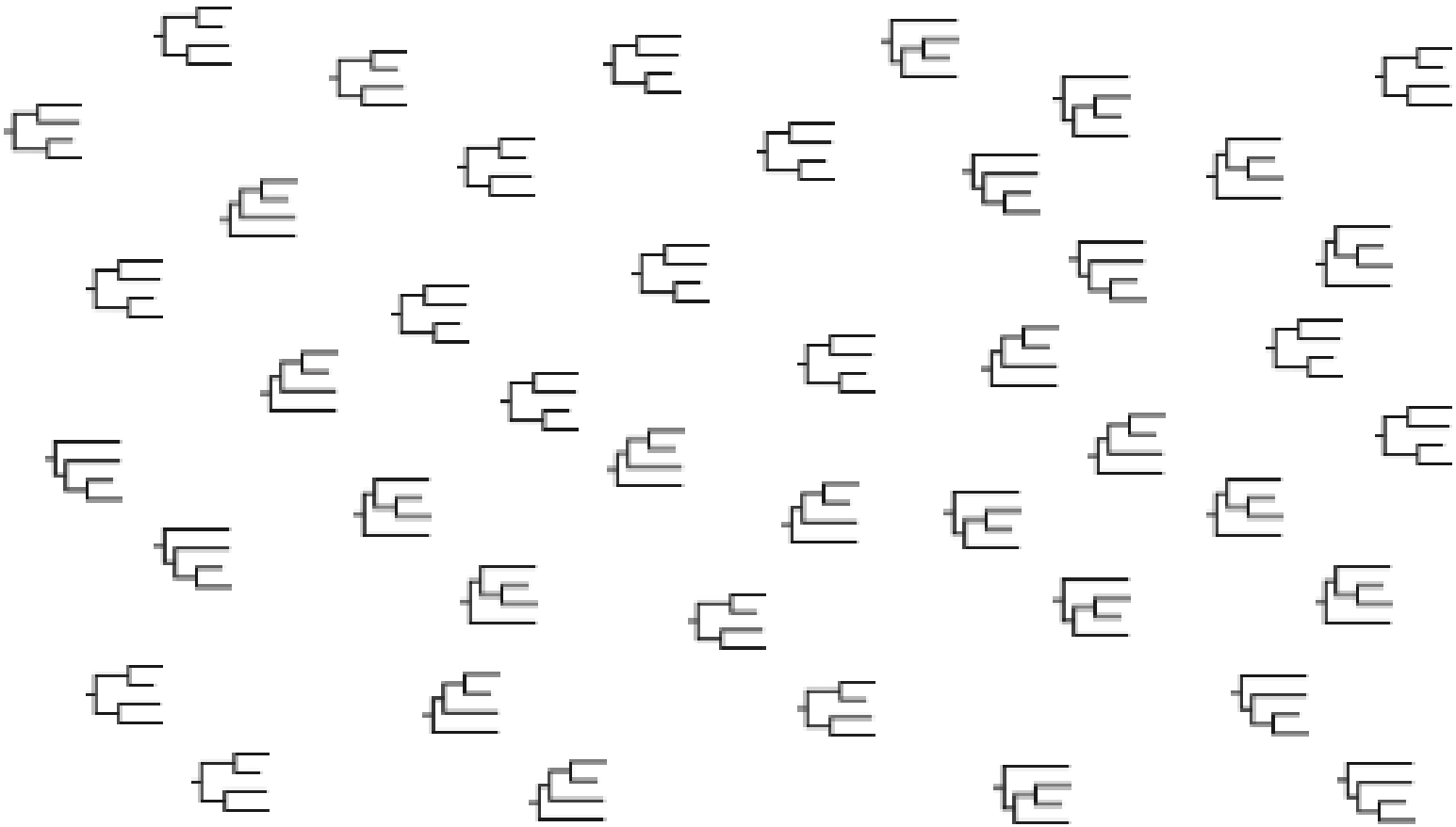
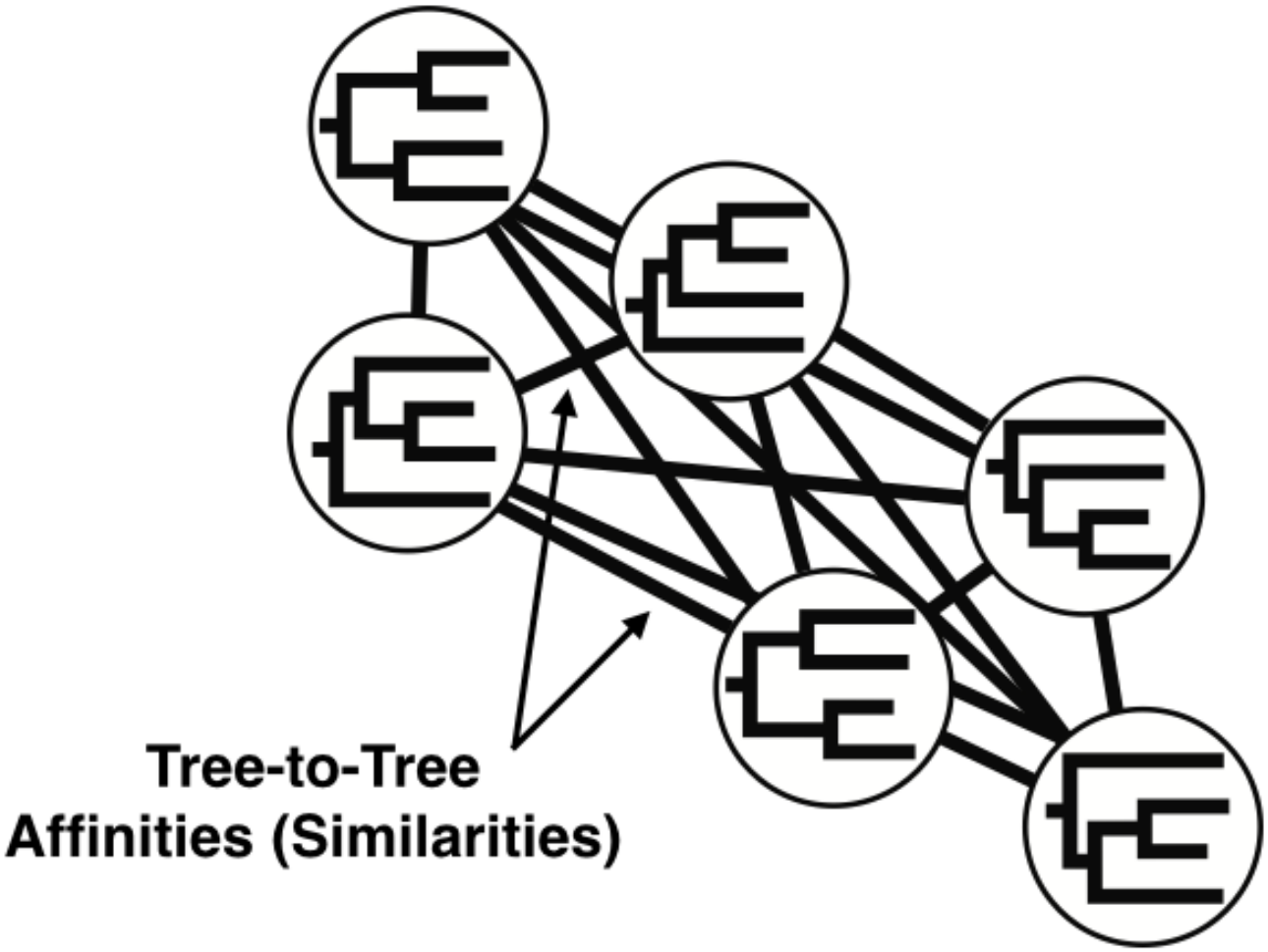


Figure: Tree Sets

Networks of Trees

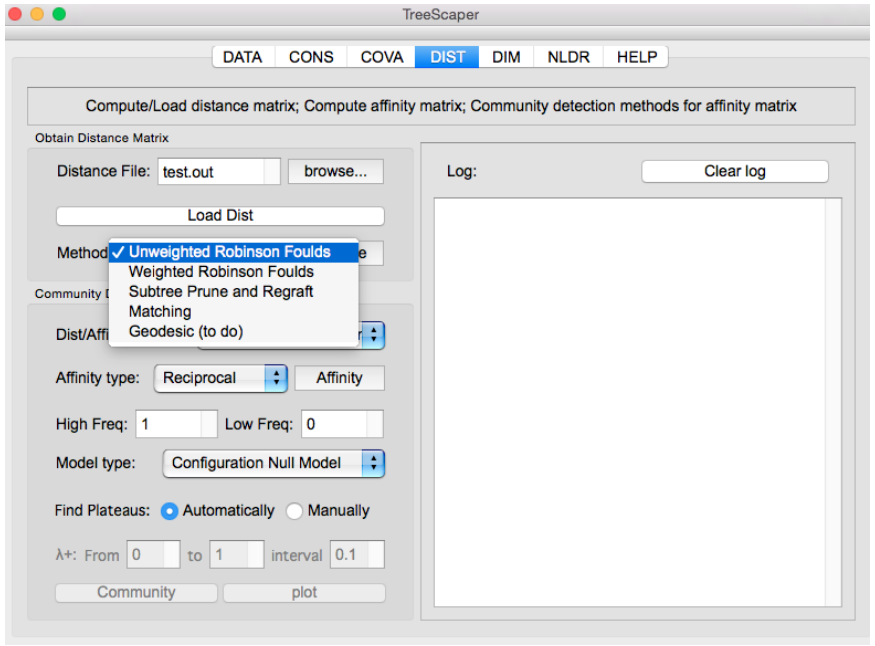
- Type I: Topology-based Network

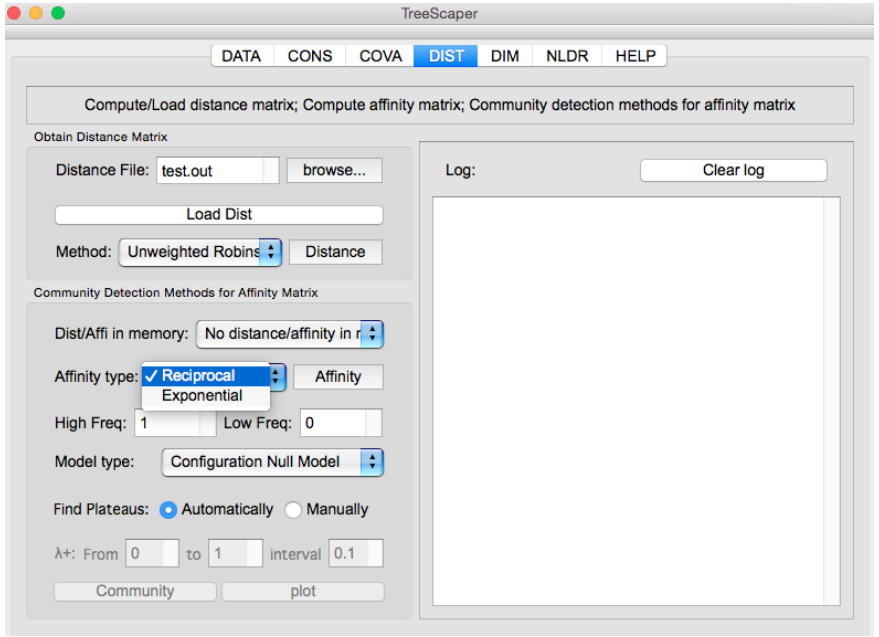
  - Nodes: trees
  - Links: topological (dis)similarities



Networks of Trees

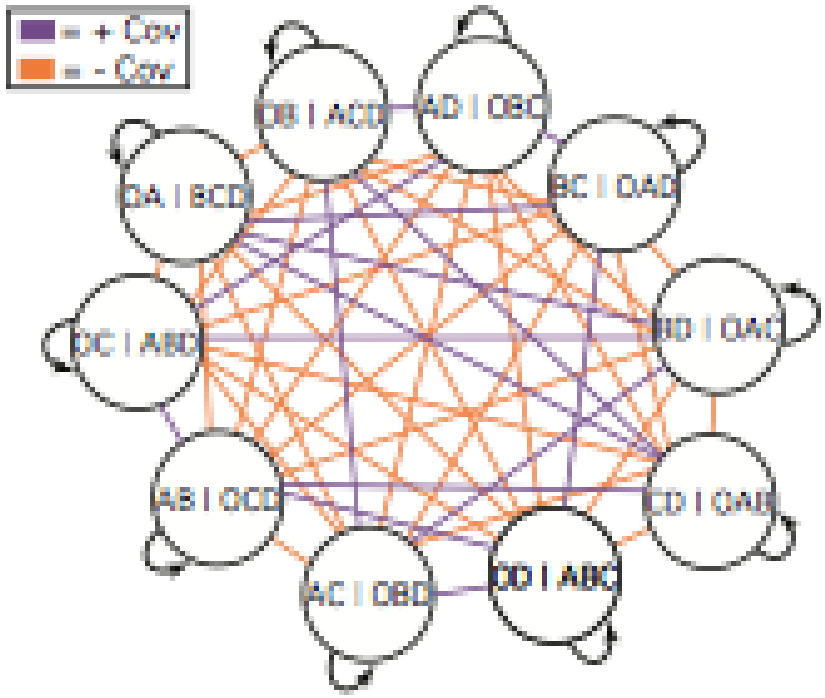
- Type I: Topology-based Network
  - Links: Affinity matrix
    - Reciprocal of pairwise distance
    - Exponential of pairwise distance





Networks of Trees

- Type II: Bipartition-based Network
  - Nodes: bipartitions
  - Links: covariance values



Community

- A **community** is a group of related nodes that
  - are densely interconnected
  - have sparser connections with the rest of the network

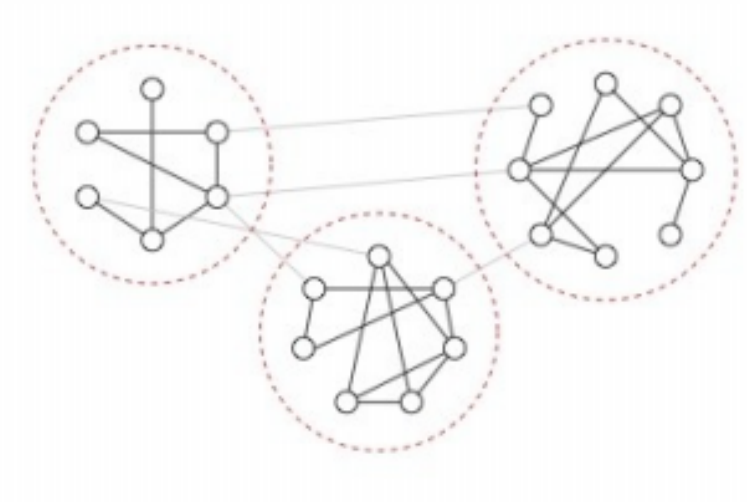
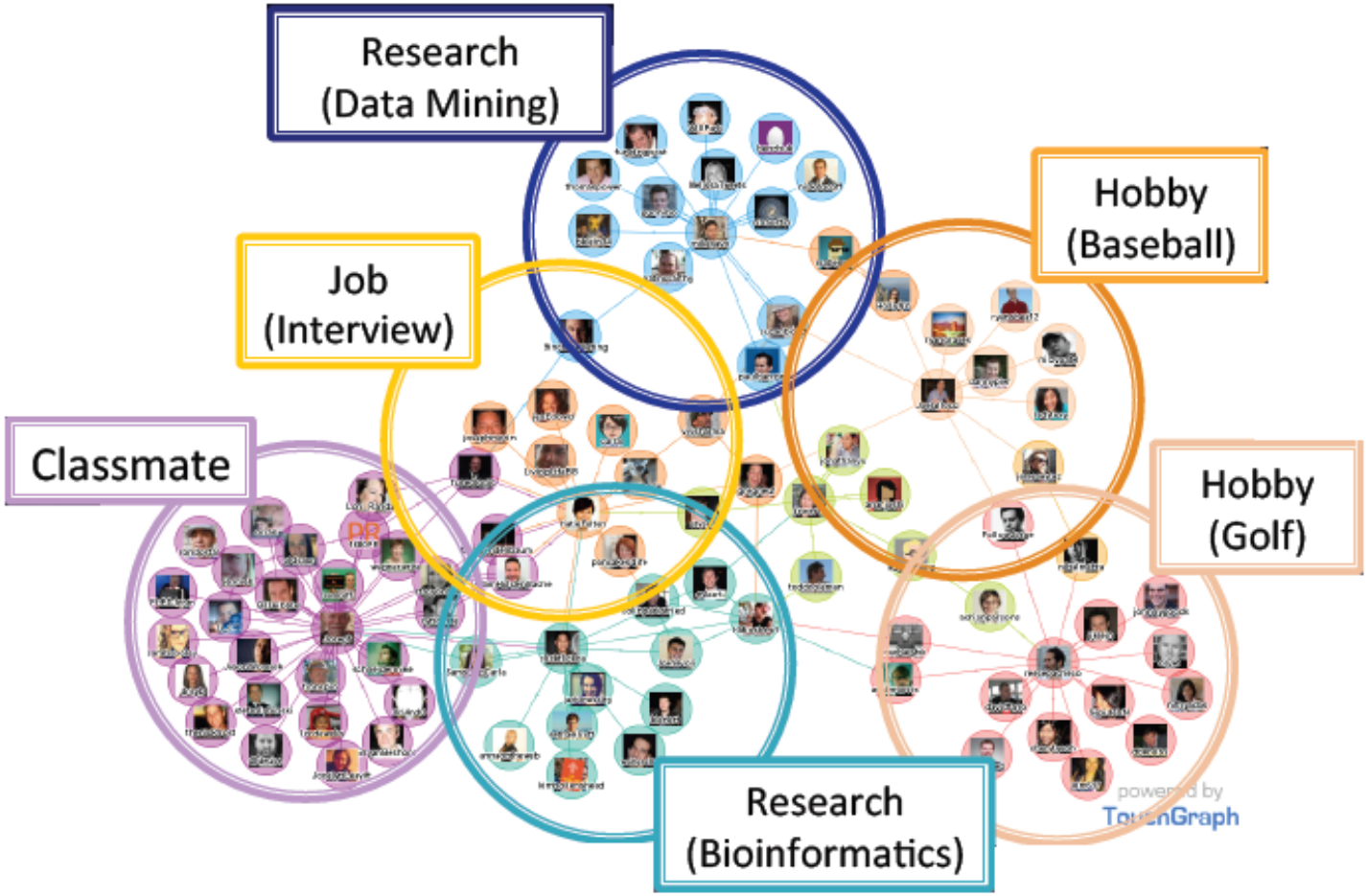


Figure: A small network with community structure

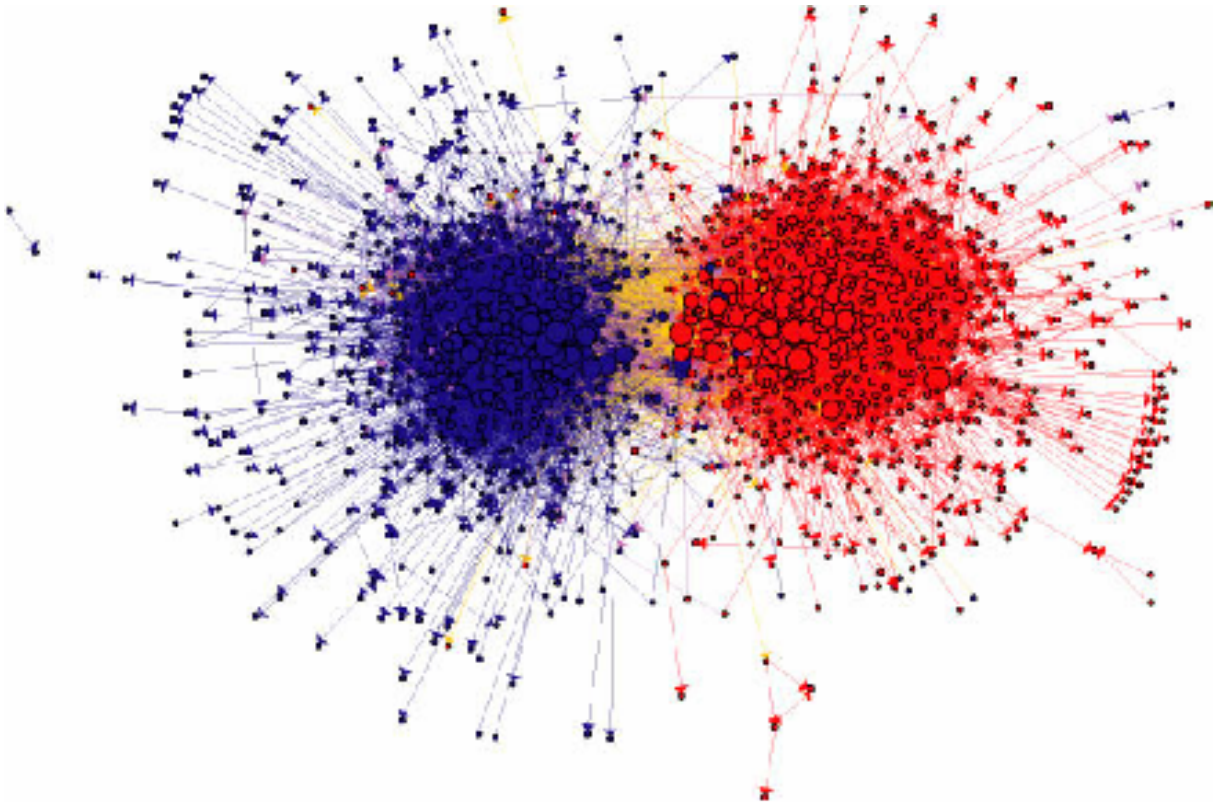
Network Communities

- Social networks



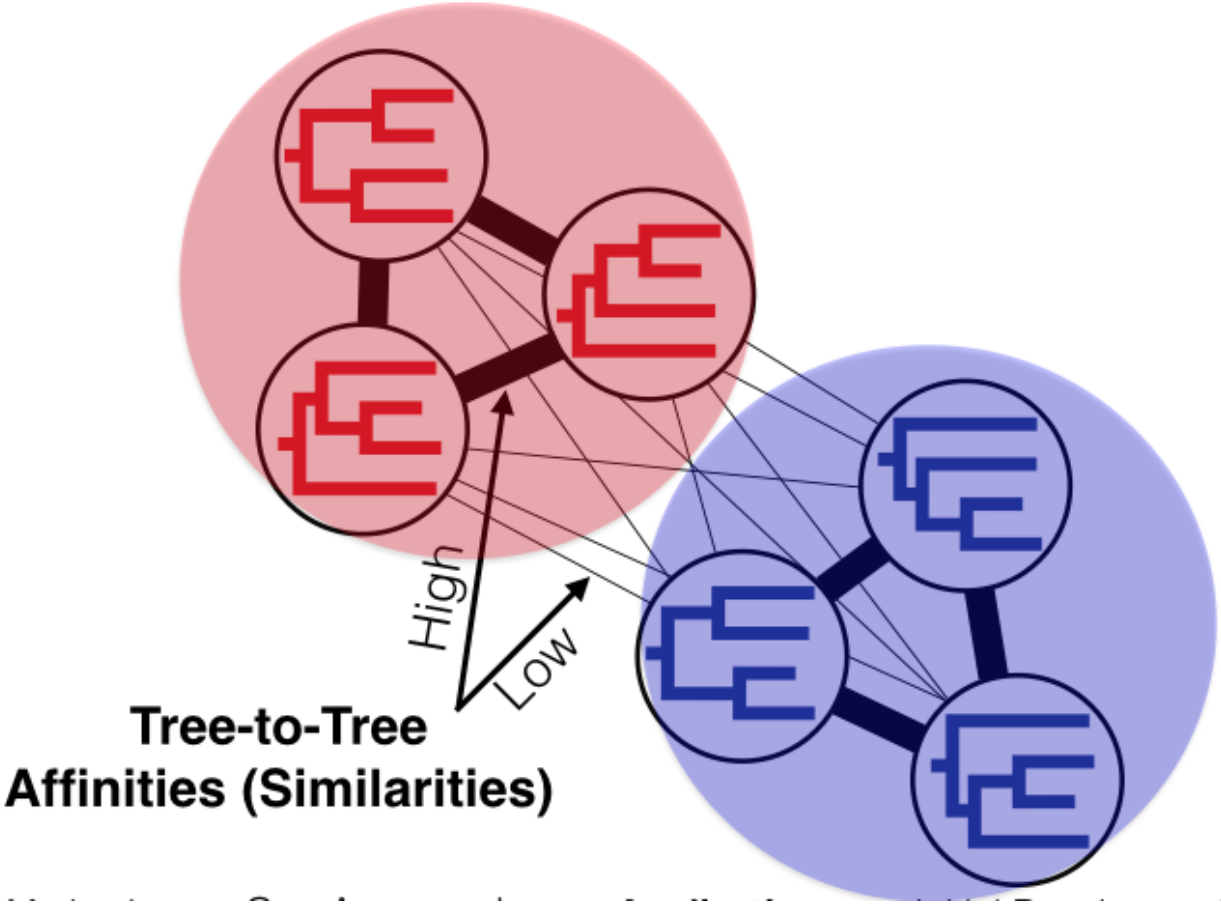
Network Communities

- Citation networks



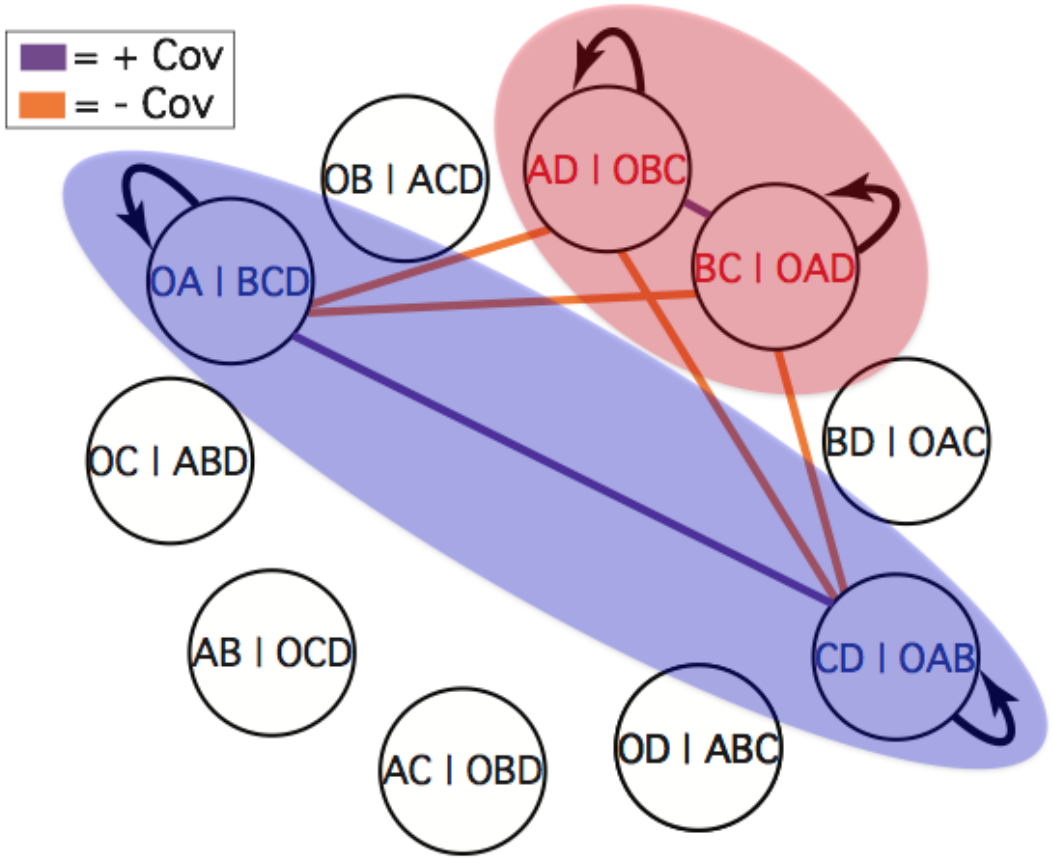
Networks of Trees

- Type I: Topology-based Network



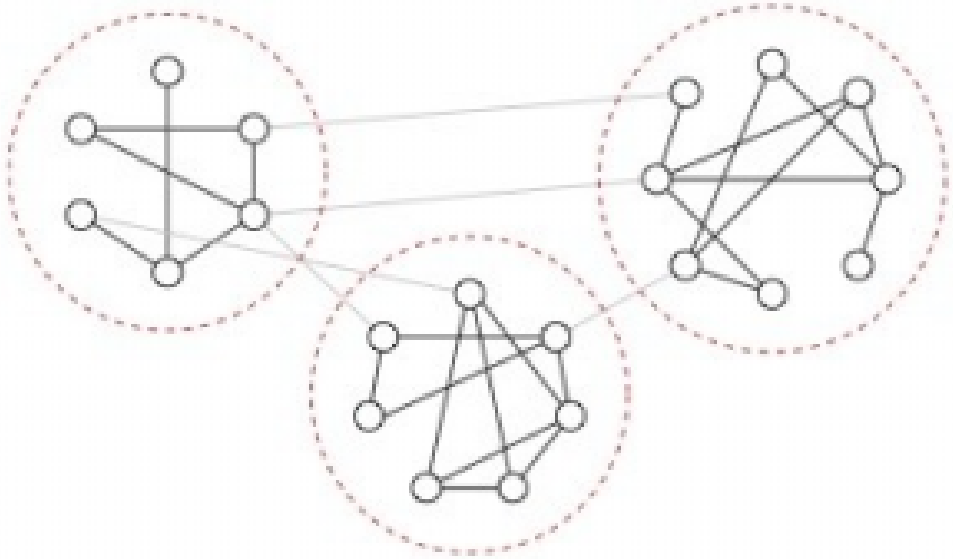
Networks of Trees

- Type II: Bipartition-based Network



Community Detection

- How can we divide the network into several parts?
  = How can we find the "community" structure?



Methods to discover communities

- No Null Model:

$$\mathcal{H}(\{\sigma\}) = -\sum_{i,j} A_{i,j}\delta(\sigma_i, \sigma_j).$$

- Compare to a randomized network:
  - Erdos-Renyi Model

$$\mathcal{H}(\{\sigma\}) = -\sum_{i,j}[A_{i,j} - c^2(\lambda^+ p_{ij}^+ - \lambda^- p_{ij}^-)]\delta(\sigma_i, \sigma_j),$$

where p_{ij} is the probability of a positive (p_{ij}^+) or negative (p_{ij}^-) between nodes i, j, λ^+, λ^- are tuning parameters.

Methods to discover communities

- Compare to a randomized network:
  - Configuration Null Model

$$\mathcal{H}(\{\sigma\}) = -\sum_{i,j}[A_{i,j} - \lambda^{+}\frac{k_i^{+}k_j^{+}}{m^{+}} - \lambda^{-}\frac{k_i^{-}k_j^{-}}{m^{-}})]\delta(\sigma_i, \sigma_j),$$

    where k_i is either the sum of the absolute value of all positive edges (k_i^{+}) or negative edges (k_i^{-}) of node i. m is either the sum of the absolute values of all positive edges (m^{+}) or negative edges (m^{-}).

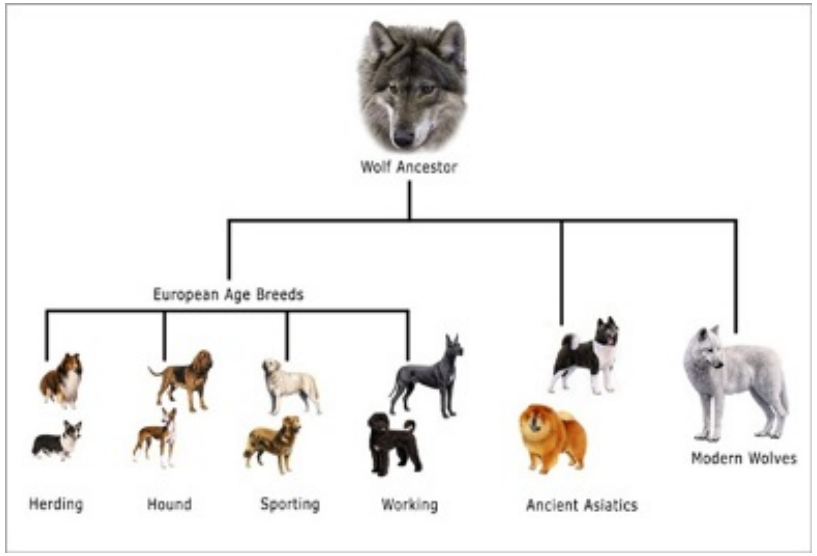
- Constant Potts Model

$$\mathcal{H}(\{\sigma\}) = -\sum_{i,j}[A_{i,j} - c^2(\lambda^{+} - \lambda^{-})]\delta(\sigma_i, \sigma_j),$$

where c is the size of the community.

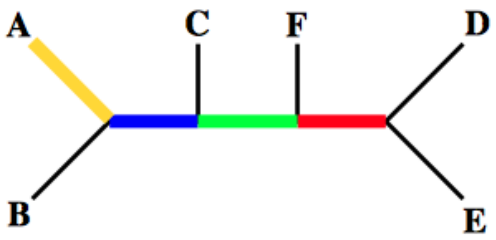
Phylogenetics

- Phylogenetic trees illustrate the evolutionary relationships among species, populations, individuals or genes (taxa in a general sense)
- The results of phylogenetic analysis are usually presented as a collection of nodes and branches.

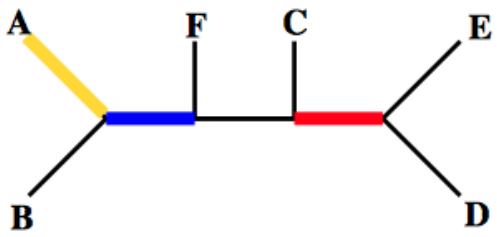


Bipartition Representation

- Evolutionary relationships are represented by edges
- Remove an edge, the taxa (nodes) will be split into two nonempty subsets
- A phylogenetic tree can be represented as a set of splits (bipartitions)



AB | CDEF
ABC | DEF
ABCF | DE



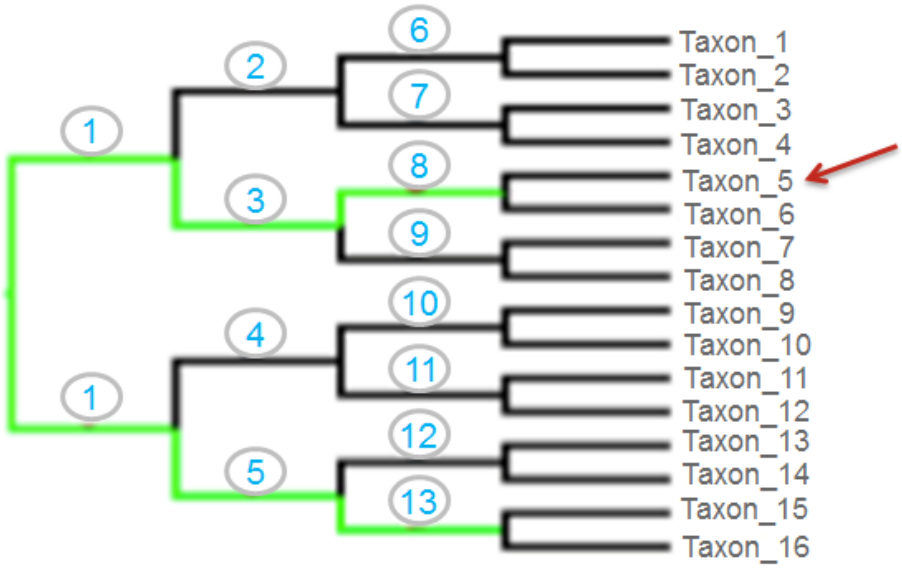
AB | CDEF
ABF | CDE
ABCF | DE

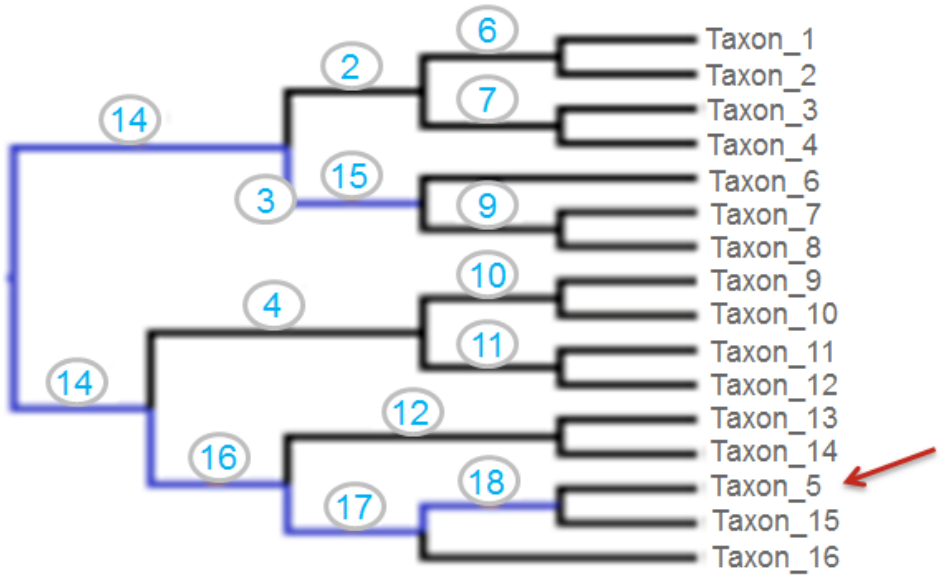
Project Motivations

- Multi-source data often produce conflicting trees
- Existing methods hide potential conflicts
  - Consensus tree
    - Discards information concerning competing trees
  - Project into low dimensional Euclidean space
    - May be difficult to interpret
- Community detection is used to explore conflicting signal in sets of phylogenies
- Develop software to analyze phylogenetic data

Method

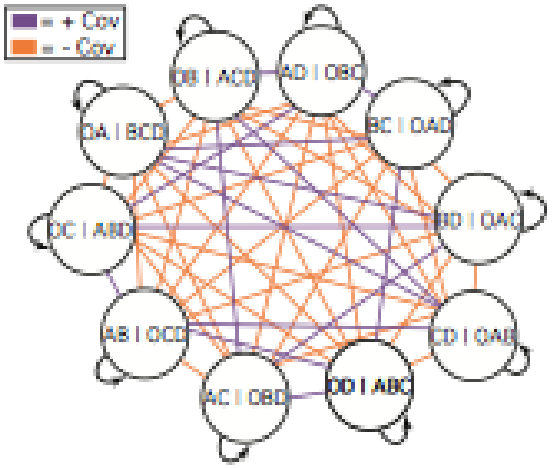
- Simulation of tree sets with conflicting signals
- Two guide trees are only differed in their placement of taxon 5 (the rogue taxon)



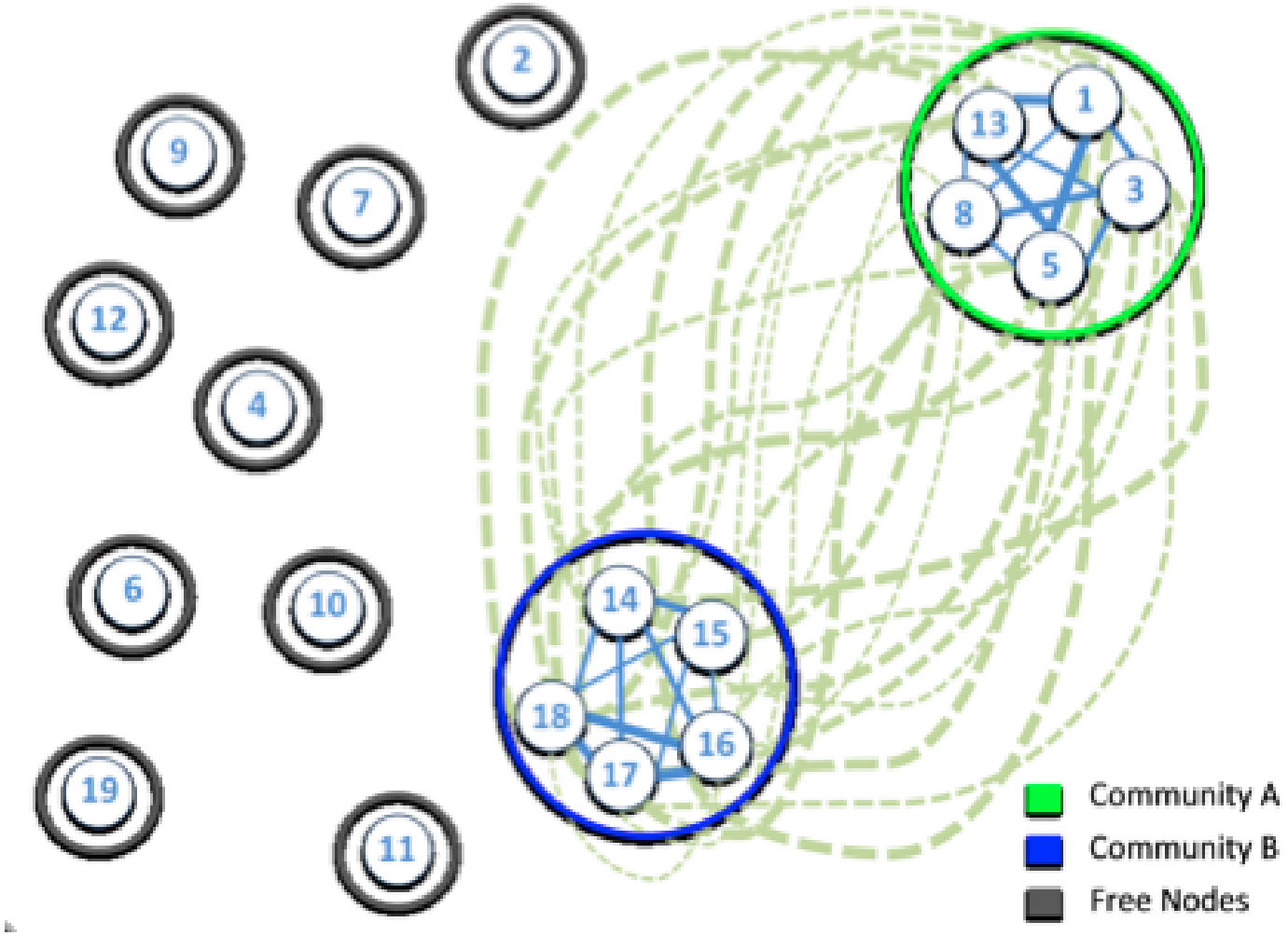


Method

- Covariance matrix based on presence or absence of bipartitions in the phylogenetic trees
- Construct a network by covariance matrix
  - Nodes: bipartitions
  - Links: covariance values



Experiment Results



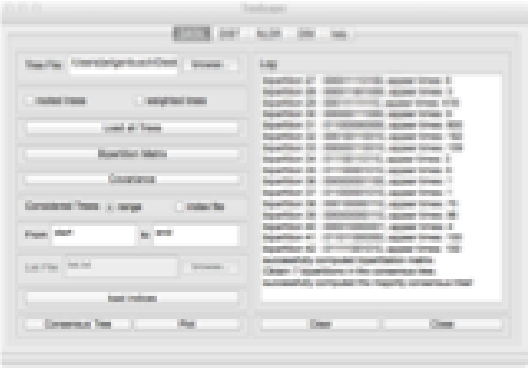
TreeScaper Software

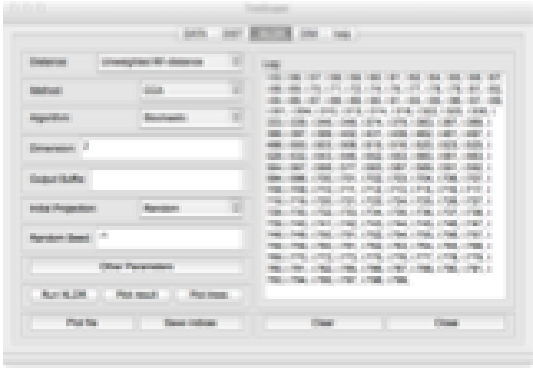
Functionality:

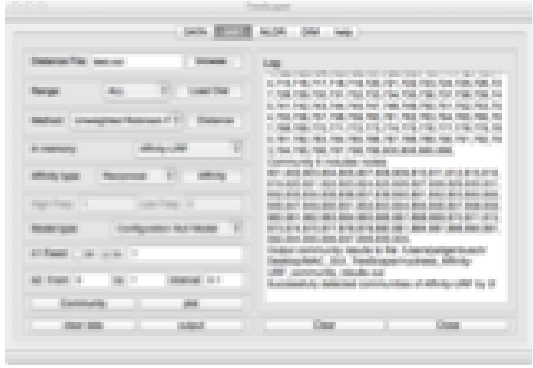
- NLDR
- Dimensionality estimation
- Distance/Affinity matrix
- Covariance matrix
- Community Detection methods
- Interactive visualization interface

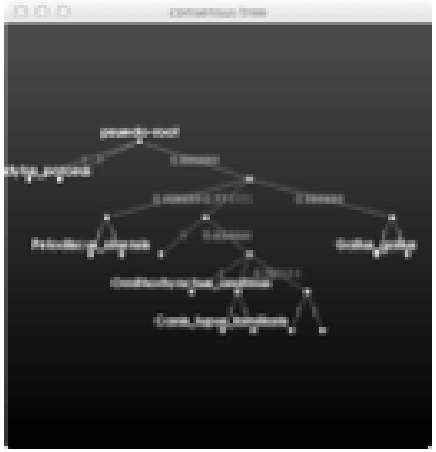
TreeScaper Software

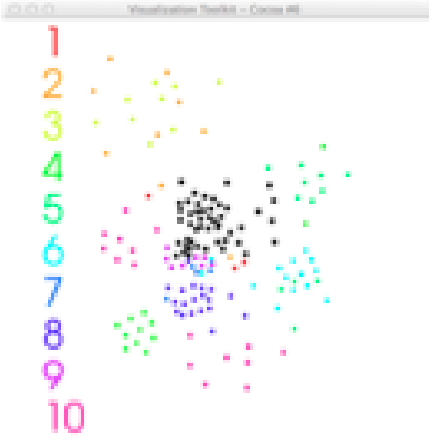
- Link: http://sourceforge.net/projects/treescaper/

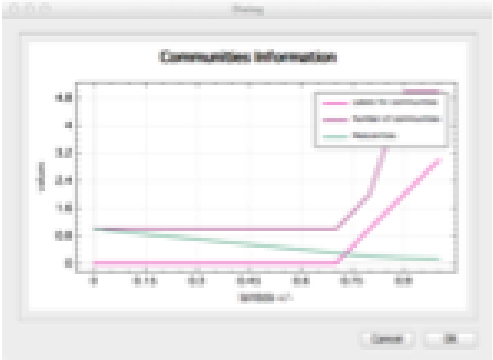












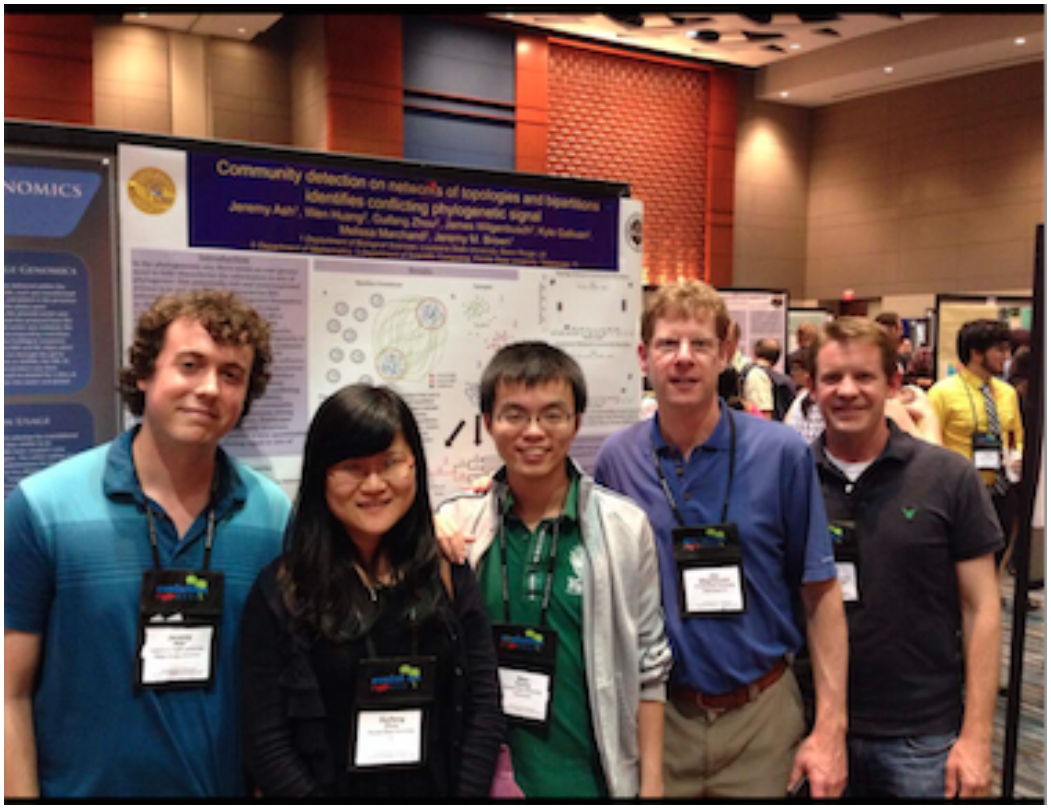
Conclusion

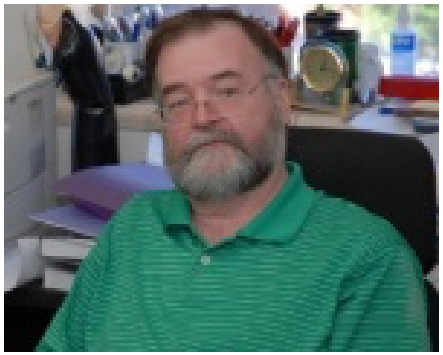
- Networks exist in many fields
- Community detection provides a valuable tool for understanding structure in massive networks
- Many methods are capable to detect communities
- Which one(s) is better? Choice depends on
  - Metric
  - Algorithm
  - Relationship to the computing platform
- Community detection provides a new quantitative approach for exploring conflicting signal in phylogenetic data.

Future Work

- Other community detection methods
  - Overlapping Community Detection
- Other network properties
  - Centrality
  - Similarity

Team Members





Acknowledgements

- FSU's Shared High Performance facility for compute cycles and technical support
- The National Science Foundation for funding to support some of this work (ABI-1262476)

THANK YOU !

- M. E. J. Newman and M. Girvan, *Finding and evaluating community structure in networks*, Phys. Rev. E **69** (2004), 026113.
- M. Barthelemy S. Fortunato, *Resolution limit in community detection*, PNAS **104** (2007), 36–41.
- V. A. Traag, P. Van Dooren, and Y. Nesterov, *Narrow scope for resolution-limit-free community detection*, Phys. Rev. E **84** (2011), 016114.
- R.Lambiotte E.Lefebvre V.E.Blondel, J.-L.Guillaume, *Fast unfolding of communities in large networks*, Journal of Statistical Mechanics: Theory and Experiment (2008), no. 10, P10008".