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

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

October 26, 2015

Outline





3 Application





æ

(日) (周) (三) (三)



What are networks?



3

(日) (周) (三) (三)

Computer Networks



Figure: A small computer network

Jeremy M. Brown , Guifang Zhou , Jeremy A Network Analysis on Phylogenetic Data

3

<ロ> (日) (日) (日) (日) (日)

Introduction

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 > < = > < = >

Jeremy M. Brown , Guifang Zhou , Jeremy A Network Analysis on Phylogenetic Data

Networks of Trees?



Figure: Tree Sets

3

<ロ> (日) (日) (日) (日) (日)

Introduction

Networks of Trees

• Type I: Topology-based Network

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



Image: A match a ma

Introduction

Networks of Trees

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

Contraction of the second seco
DATA CONS COVA DIST DIM NLDR HELP
Compute/Load distance matrix; Compute affinity matrix; Community detection methods for affinity matrix
Obtain Distance Matrix
Distance File: test.out browse Log: Clear log
Load Dist
Method: Unweighted Robins : Distance
Community Detection Methods for Affinity Matrix
Dist/Affi in memory: No distance/affinity in r
Affinity type: Reciprocal Affinity
High Freq: 1 Low Freq: 0
Model type: Configuration Null Model
Find Plateaus: O Automatically O Manually
λ +: From 0 to 1 interval 0.1

Image: A match a ma

Networks of Trees

• Type II: Bipartition-based Network

- Nodes: bipartitions
- Links: covariance values



3

< ロ > < 同 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ >

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



- 32

イロン イヨン イヨン イヨン

Networks of Trees

• Type I: Topology-based Network



∃ →

Image: A math a math

Networks of Trees

• Type II: Bipartition-based Network



3

(日) (周) (三) (三)

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 $i, j, \lambda^+, \lambda^-$ are tuning parameters.

- 4回 ト 4 ヨ ト - 4 ヨ ト - - ヨ

Methods to discover communities

- Compare to a randomized network:
 - Configuration Null Model

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

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.



A B A B A B A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A
 A

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)



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



25 / 32

TreeScaper Software

Functionality:

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

3

< 🗗 🕨 🔸

TreeScaper Software

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

barta tertifythetitet aven	ing .
Contrast Congrations	Apartice in conversion again from 1 Apartice 2 (Section 1), again from 12
Load at Team	Apartment of Collegence and Street Street Street
Rearting Maria	Spatter 5 Control and the 19
Georgean	Applie & CONTRACT, applied in
Content Test () range () Contention	Apartica in concentral again that it
ton Art is an	Apartics in conversion wave true 1
and the local division of	Apartitar & Provident Apartment Inter- Oper 7 spartition in the american line.
had indian	

helerosi.	invegine W-deares		-
adval .	414		************
age that	Barbarb		10.00.00.00.00.00.00.00.00.00.00.00.00.0
inanan i			48. 00. 00. 00. 00. 01. 00. 00. 00. 00. 00. 00. 01. 00. 01. 00. 00. 00. 00.
lege lafe			
ita Popula	a Anton	1	14-14-18-19-18-19-18-19-19-19-1 18-18-19-19-19-19-19-19-19-19-19-1
landor (basi			
(the Parameters			
Aur 16,218), Patron (, Patro	-	

		1.00
tege (as)	town in	
Mark International	() (march)	
to rearrange and the second	an 1	1 To 10,
day yes . Annual	1 mg	Nor- and all states and all states are an an art and all states are
spring 1 cards		No. 11, 10, 10, 10, 10, 10, 10, 10, 10, 10,
And the second	name 1	111214-011-011-011-00-00-00-00-00-00-00-00-
1. Aug		Contraction of the Contraction o
1 Tan 1 Tan 1	manal are	UNF community, sticule and Reconnecting deleted communities of Allering-pilding in
Same I	-	







< 一型

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.

イロト 不得下 イヨト イヨト

Future Work

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

3

(日) (同) (三) (三)

Team Members



Jeremy M. Brown , Guifang Zhou , Jeremy A Network Analysis on Phylogenetic Data

イロト イ団ト イヨト イヨト 二日

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)

イロト 不得下 イヨト イヨト

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".

イロト 不得下 イヨト イヨト