Network Phylogenies of Abui

An initial look at inferring language contact

Gereon A. Kaiping

Leiden University Centre for Linguistics, Niederlande

2019-08-21

1. Network Phylogenies
2. Abui
3. Results
4. Discussion

Why infer phylogenetic networks?

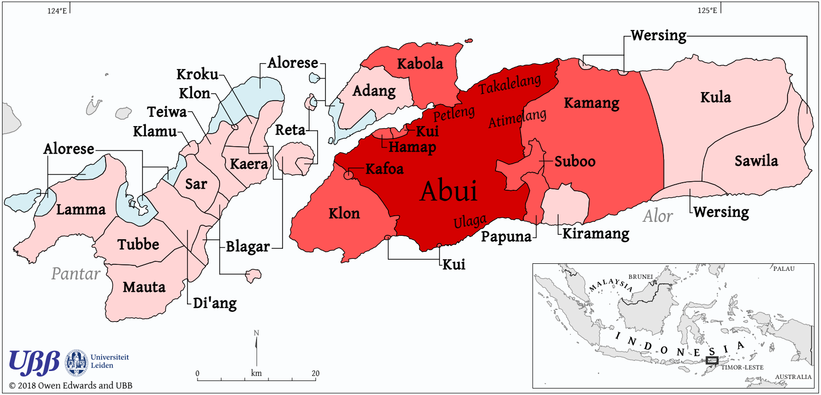
- Language contact: important driver of language evolution
- Tree assumption in phylogenetics is limiting
    - Dealing with borrowing as pre-cleanup is hard
    - Language contact is part of the history to be inferred
- Grouping word trees for understanding strata in the vocabulary

State of the art

- Big bubble of Bayesian phylogenetics in linguistics, with slowly improving tools
- SplitsTrees, NeighborNets → No model, only visualization
- Initial bits of Bayesian network inference in genetics

Here: A (first?) practical look at network inference for languages

Language sample: Abui



Recipe

Full analysis available under
`http://github.com/Anaphory/abui-network`

- BEASTling[1] configuration with rate variation
  - Abui & Neighbors
  - Data from LexiRumah[2] with ACD[3]
  - pseudo-Dollo Covarion model[4]
- Python script to add Species Network[5], grouping ‘gene’ trees
- Phylogenetic inference in BEAST[6]

[1] Maurits et al. 2017.

[2] Kaiping & Klamer 2018.

[3] List 2012.

[4] Bouckaert & Robbeets 2017.

[5] Zhang et al. 2018.

[6] Bouckaert et al. 2014.

Results

Practical issues

- SpeciesNetwork was buggy
- SN takes only small data (genes, taxa)
- Driver file construction is complicated
- Displaying results

Obvious issues

- How to summarize results? (!)
- How many trees to infer?
- How to improve the MCMC?

Theoretical issues

- Multispecies Network Coalescent prior: good?
- Population model?
- Convergence vs. local maximum?
- Non-lexical data?
- Calibrations?
- What amount of reticulation should be expect?
- Do we have extremely tree-like histories to test this on?

Results Issues

Practical issues

- SpeciesNetwork was buggy
- SN takes only small data (genes, taxa)
- Driver file construction is complicated
- Displaying results

Obvious issues

- How to summarize results? (!)
- How many trees to infer?
- How to improve the MCMC?

Theoretical issues

- Multispecies Network Coalescent prior: good?
- Population model?
- Convergence vs. local maximum?
- Non-lexical data?
- Calibrations?
- What amount of reticulation should be expect?
- Do we have extremely tree-like histories to test this on?

ResultsIssues

Practical issues

- SpeciesNetwork was buggy
- SN takes only small data (genes, taxa)
- Driver file construction is complicated
- Displaying results

Obvious issues

- How to summarize results? (!)
- How many trees to infer?
- How to improve the MCMC?

Theoretical issues

- Multispecies Network Coalescent prior: good?
- Population model?
- Convergence vs. local maximum?
- Non-lexical data?
- Calibrations?
- What amount of reticulation should be expect?
- Do we have extremely tree-like histories to test this on?

ResultsIssues

Practical issues

- SpeciesNetwork was buggy
- SN takes only small data (genes, taxa)
- Driver file construction is complicated
- Displaying results

Obvious issues

- How to summarize results? (!)
- How many trees to infer?
- How to improve the MCMC?

Theoretical issues

- Multispecies Network Coalescent prior: good?
- Population model?
- Convergence vs. local maximum?
- Non-lexical data?
- Calibrations?
- What amount of reticulation should be expect?
- Do we have extremely tree-like histories to test this on?

ResultsIssues

Practical issues

- SpeciesNetwork was buggy
- SN takes only small data (genes, taxa)
- Driver file construction is complicated
- Displaying results

Obvious issues

- How to summarize results? (!)
- How many trees to infer?
- How to improve the MCMC?

Theoretical issues

- Multispecies Network Coalescent prior: good?
- Population model?
- Convergence vs. local maximum?
- Non-lexical data?
- Calibrations?
- What amount of reticulation should be expect?
- Do we have extremely tree-like histories to test this on?

ResultsIssues

Practical issues

- SpeciesNetwork was buggy
- SN takes only small data (genes, taxa)
- Driver file construction is complicated
- Displaying results

Obvious issues

- How to summarize results? (!)
- How many trees to infer?
- How to improve the MCMC?

Theoretical issues

- Multispecies Network Coalescent prior: good?
- Population model?
- Convergence vs. local maximum?
- Non-lexical data?
- Calibrations?
- What amount of reticulation should be expect?
- Do we have extremely tree-like histories to test this on?

ResultsIssues

Practical issues

- SpeciesNetwork was buggy
- SN takes only small data (genes, taxa)
- Driver file construction is complicated
- Displaying results

Obvious issues

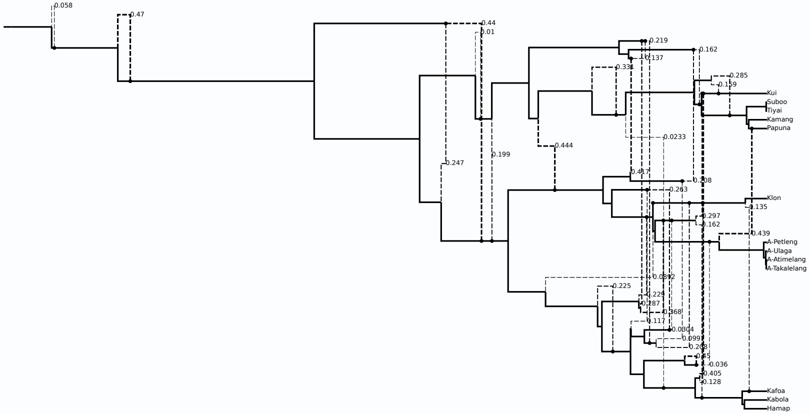
- How to summarize results? (!)
- How many trees to infer?
- How to improve the MCMC?

Theoretical issues

- Multispecies Network Coalescent prior: good?
- Population model?
- Convergence vs. local maximum?
- Non-lexical data?
- Calibrations?
- What amount of reticulation should be expect?
- Do we have extremely tree-like histories to test this on?

Results

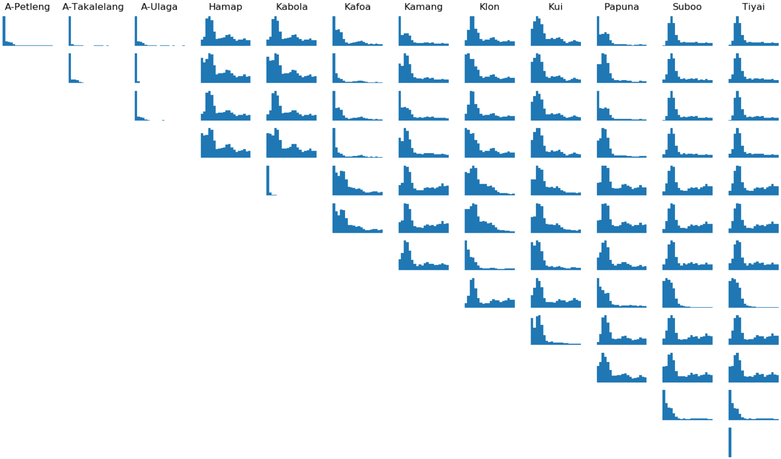
A typical (?) network[7]



[7] https://icytree.org/

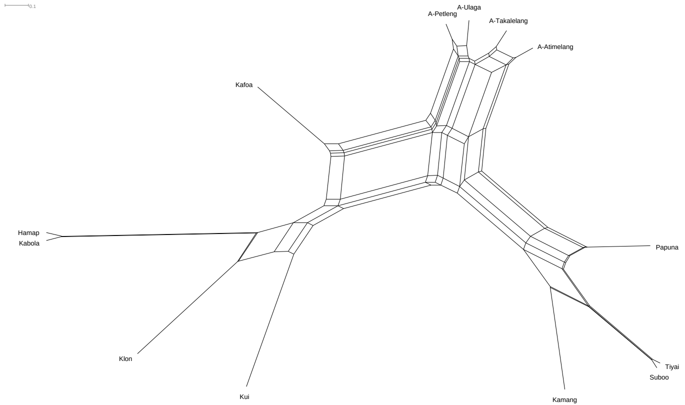
Results

Summary of pairwise distances



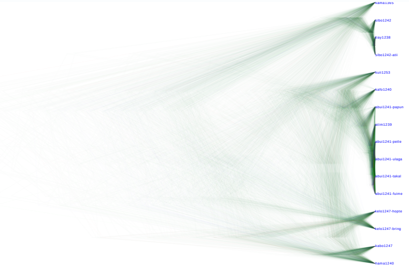
Results

SplitsNetwork from mean distances

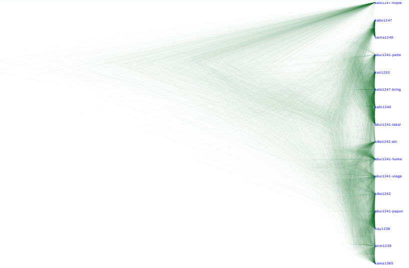


Computer-*assisted*?

Words that most often follow different trees



dark, water, to-stand, wing, to-search-for, fog, sweet, comb, nose, woman, to-spit, to-die, bad, star, fingernail



ripe, twenty-one, to-smell, fire, twenty, twelve, thirteen, to-buy, thirty, oven, eleven, sugar-palm, 3pl, horn, seventy

Candidates for informative vs. noisy concepts?

Summary

- Abui & neighbours (Kafoa, Papuna!) are inferred with a lot of contact signal
- No a-priori borrow detection, maybe even find strata in the lexicon
- Start adding networks to our toolbox, solve outstanding issues
  - Modeling
  - Technology
  - Validation
  - Visualization

`http://github.com/Anaphory/abui-network`

References I

- Bouckaert, Remco & Heled, Joseph & Kühnert, Denise & Vaughan, Tim & Wu, Chieh-Hsi & Xie, Dong & Suchard, Marc A. & Rambaut, Andrew & Drummond, Alexei J. 2014. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput Biol* 10(4). e1003537. https://doi.org/10.1371/journal.pcbi.1003537.
- Bouckaert, Remco & Robbeets, Martine. 2017. Pseudo Dollo models for the evolution of binary characters along a tree. *bioRxiv*. 207571. https://doi.org/10.1101/207571.
- Kaiping, Gereon A. & Klamer, Marian. 2018. LexiRumah: An online lexical database of the Lesser Sunda Islands. *PLOS ONE* 13(10). e0205250. https://doi.org/10.1371/journal.pone.0205250.

References II

- List, Johann-Mattis. 2012. LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH* (EACL 2012), 117–125. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Maurits, Luke & Forkel, Robert & Kaiping, Gereon A. & Atkinson, Quentin D. 2017. BEASTling: A software tool for linguistic phylogenetics using BEAST 2. *PLOS ONE* 12(8). e0180908. `https://doi.org/10.1371/journal.pone.0180908`.
- Schleicher, August. 1871. *Compendium der vergleichenden Grammatik der indogermanischen Sprachen*. Böhlau. 894 pp.

References III

Zhang, Chi & Ogilvie, Huw A. & Drummond, Alexei J. & Stadler, Tanja. 2018. Bayesian Inference of Species Networks from Multilocus Sequence Data. *Molecular Biology and Evolution* 35(2). 504–517. `https://doi.org/10.1093/molbev/msx307`.