Quartet concordance factors for networks beyond level-1

To get our working group up and running quickly, this document contains some suggested readings and a homework assignment to be completed before we arrive in Hawaii. As group members have different knowledge bases and are at different points in their careers, do not worry if some of this reading is a stretch. You should, however, get at least the 'sense' of things listed explicitly below.

Please recall that our project goals are: Working under the NMSC on a binary semi-directed network to

(1) Understand how the NMSC on an arbitrary network leads to a polynomial form for quartet CFs, and work these out for small networks (e.g., 6-taxon) with specific non-level-1 structures (e.g., level-2 blobs),

(2) Determine the resulting varieties for these networks, with a goal of understanding their intersections and containments,

(3) Use these results to characterize a class of networks for which identifiability by quartet CFs holds for generic parameters.

Reading list:

- For background on computing quartet concordance factors under the NMSC, please look at [4] and Section 5 in particular. Therein you will find some explicit examples of computing CFs for 4-taxon level-1 networks. In addition, please skim the paper [2] and its appendices. This paper is technical, but it has the flavor of using an algebraic approach with identifiability problems under the NMSC.

- For background on the technical definition of level-2 networks and generators, the beginning of the paper *The Structure of Level-k Phylogenetic Networks* by P. Gambette, V. Berry, and C. Paul gives the definitions and generators of level-2 networks. (Many other references are adequate too, including [5] below.) In short, you need to know what level-2 means in terms of a network graph.

- *Many* other references are excellent, and there is not really a reference quite adequate for what we need. You might want to consult any of [7, 1, 5] for additional terminology. For a technical, but good, background on the coalescent model on a species tree there is also [3]

Homework Assignment: Compute the formulas for a $CF_{abcd} = (CF_{ab|cd}, CF_{ac|bd}, CF_{ad|bc})$ for the 4-taxon binary networks containing exactly 1 cycle C which is a i) 2_2-cycle, ii) 3_1-cycle, iii) 3_2-cycle, and iv) 4_1-cycle. For checking your answers, these formulas are given in [6] and [4, Examples 2-5].

References

[1] E.S. Allman, H. Baños, and J.A. Rhodes. NANUQ: a method for inferring species networks from gene trees under the coalescent model. *Algorithms for Molecular Biology*, 14(1):24, 2019.

[2] E.S. Allman, H. Baños, M. Garrote-Lopez, and J.A. Rhodes. Identifiability of level-1 species networks from gene tree quartets, 2024.

[3] E.S. Allman, J.H. Degnan, and J.A. Rhodes. Identifying the rooted species tree from the distribution of unrooted gene trees under the coalescent. *Journal of Mathematical Biology*, 62(6):833–862, 2011.

[4] H. Baños. Identifying species network features from gene tree quartets. *Bulletin of Mathematical Biology*, 81:494–534, 2019.

[5] Sungsik Kong, Joan Carles Pons, Laura Kubatko, and Kristina Wicke. Classes of explicit phylogenetic networks and their biological and mathematical significance. 84(6):Paper No. 47, 44, 2022.

[6] C. Solís-Lemus and C. Ané. Supplementary material: Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genetics*, 12(3), 2016.

[7] Claudia Solís-Lemus and Cécile Ané. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLOS Genetics*, 12(3):1–21, 03 2016.