Full of beans:
a study on the alignment of two flowering plants classification systems

Yi-Yun Cheng and Bertram Ludaescher

{yiyunyc2,ludaesch}@illinois.edu

https://github.com/yiyunyc2/NKOS18

NKOS2018, Sept. 13, Porto, Portugal

ILLINOIS | School of
Information Sciences
The iSchool at Illinois

Reasons to align different classifications

- Difficulty to organize information using a stable unitary classification scheme over time

- In biodiversity domain, It is common for taxonomists to contradict each other's or even their own previous taxonomies

- KOS are dynamic, time-specific, and responsive to both empirical signals and human classification interests

Cheng

School of
Information Sciences
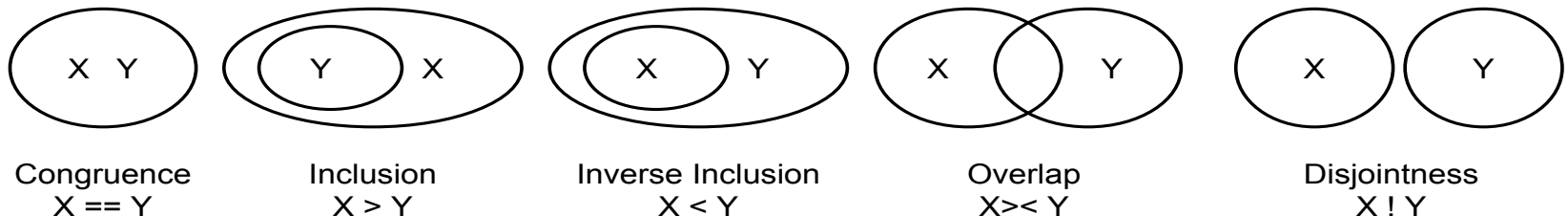The iSchool at Illinois

Purpose of this study

- Demonstrate the feasibility of integrating two classifications that result in numerous possible solutions

  – the computational power that can aid us in aligning KOS which could not have been possible when working with alignments manually

- With the hope that this work will further shed lights on:

  – the possible alignments of the classifications in the information science community

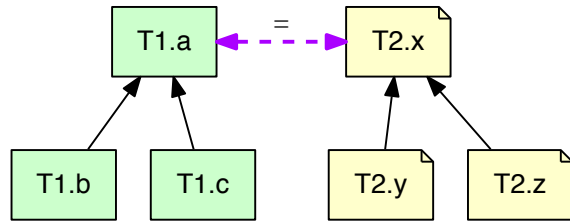 – bring a novel approach for aligning KOS in the future

School of
Information Sciences
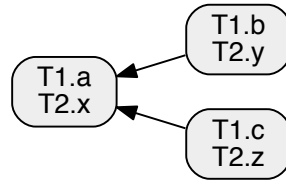The iSchool at Illinois

5 ways to relate concepts (regions)
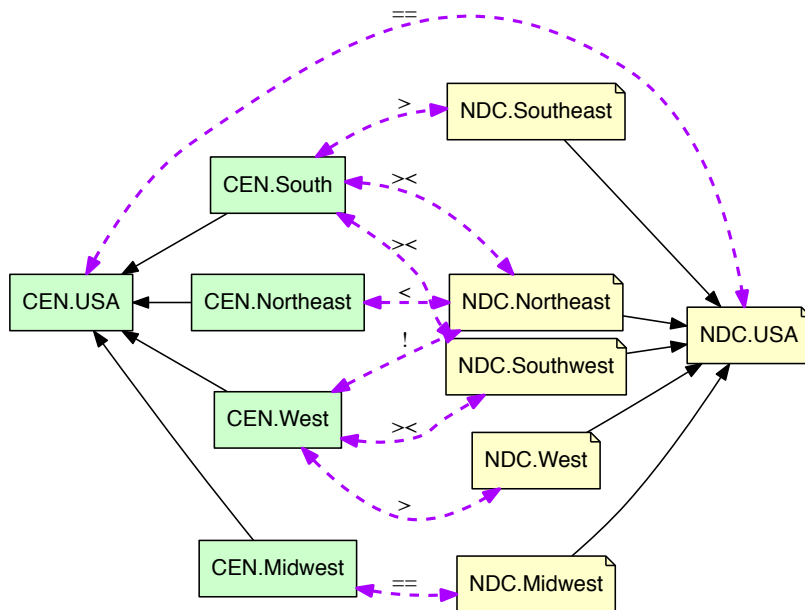
- **Idea**: relate concepts X and Y with *articulations*

- Articulation Language: **Region Connection Calculus** (**RCC5**): *congruence*, *inclusion*, *inverse inclusion*, *overlap*, *disjointness*
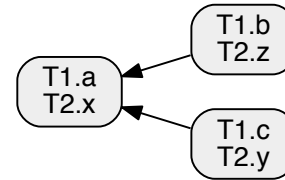
| Congruence X == Y | Inclusion X > Y | Inverse Inclusion X < Y | Overlap X>< Y | Disjointness X ! Y |

School of
Information Sciences
The iSchool at Illinois

Reasoning about taxonomies

Input Taxonomies

PW1

PW2

PW3

PW4

PW5

PW6

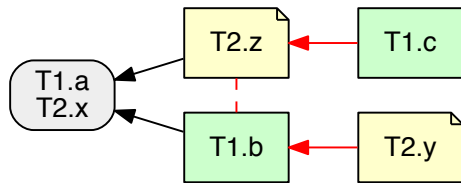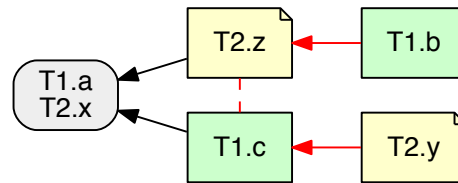*PW7*

- **Given**:
  - taxonomies T1, T2
  - and relations T1 ~ T2 (*articulations*, *alignment*)
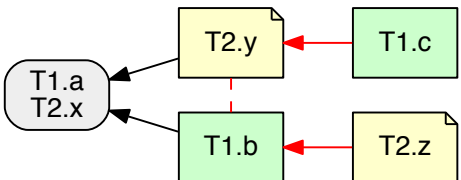- **Find**:
  - merged taxonomy T3
- **Such that**:
  - T1, T2 are **preserved**
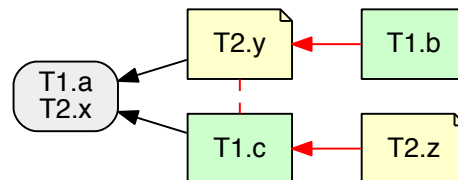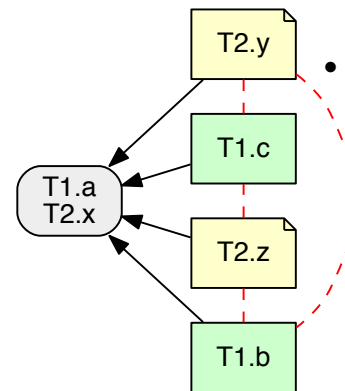 - all pairwise relations are **explicit**

School of
Information Sciences
The iSchool at Illinois

Two Flowering Plant Classifications

- Cronquist system
    - Arthur Cronquist (1981)
    - classifying plant resemblances based on evolution and *morphological* similarity (similar characteristics)
    - the most fully developed phyletic system of flowering plant classification systems by far

- Angiosperm Phylogeny Group System (APG IV)
    - uses both morphological and *molecular data* to group plants
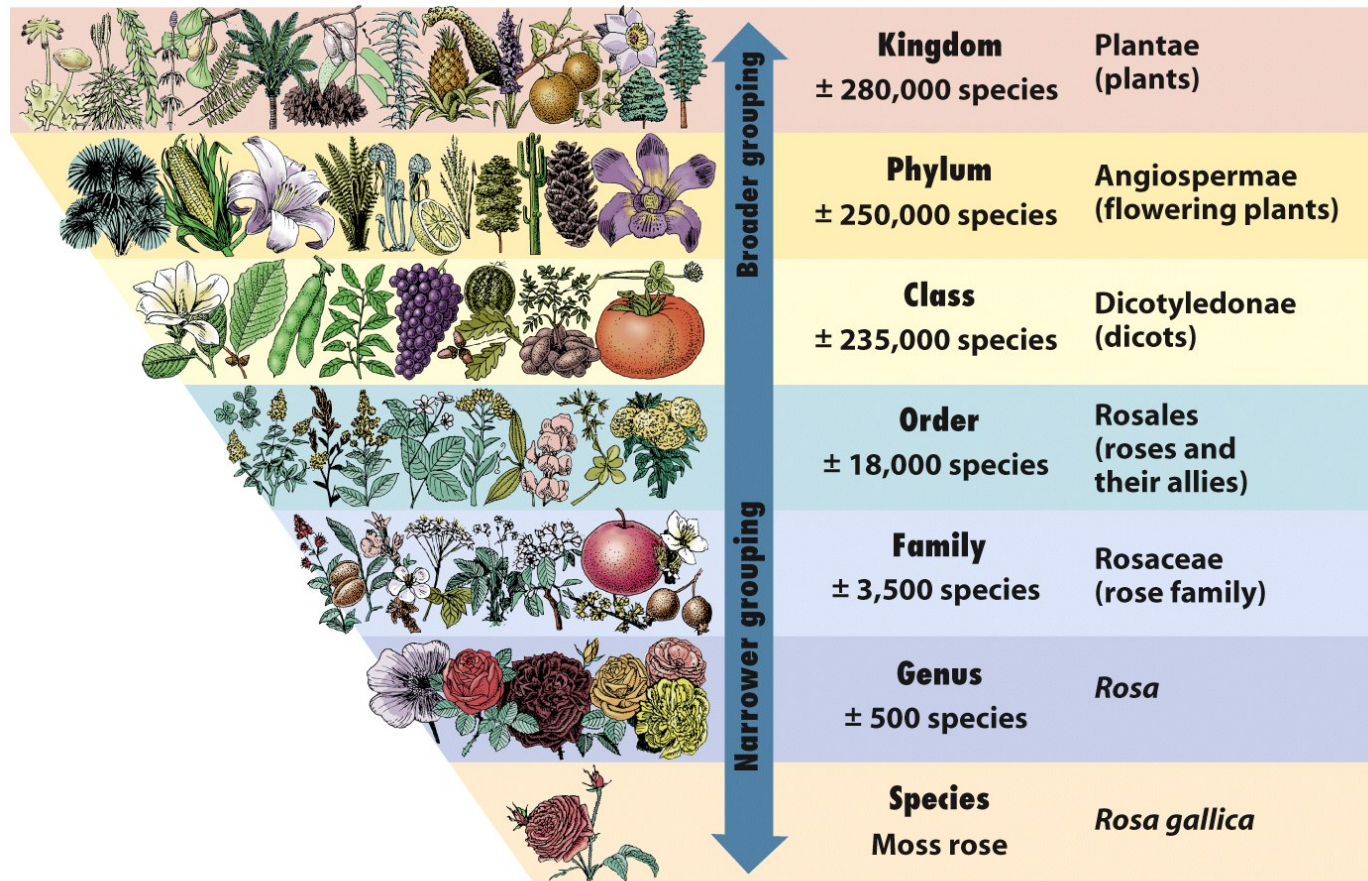   - The de facto standard classifications now

Cheng

School of
Information Sciences
The iSchool at Illinois

Biodiversity Taxonomies

Figure 2-6 Discover Biology 3/e
© 2006 W. W. Norton & Company, Inc.

School of
Information Sciences
The iSchool at Illinois

What are we aligning in our use case

- *Family* level concepts
- Total number of flowering plant families: 416
- Number of families we align
  - 40 most common families/ subfamilies
- Examples: Magnoliaceae, Ranunculaceae, Papaveraceae, Cataceae, Betulaceae, Fabaceae, Rosaceae

Cheng

School of
Information Sciences
The iSchool at Illinois

First round of alignment

- Goal:
  - Align the 40 families (or the variations of the names) in both classifications

- Rules:
  - if the family in both systems shares the **exact same name**, we assume (possible incorrectly) that they are **congruent** to each other
  - If there are **similar but different** names, we will leave the concepts unmapped at first

Cheng

School of
Information Sciences
The iSchool at Illinois

9/13/18   Cheng
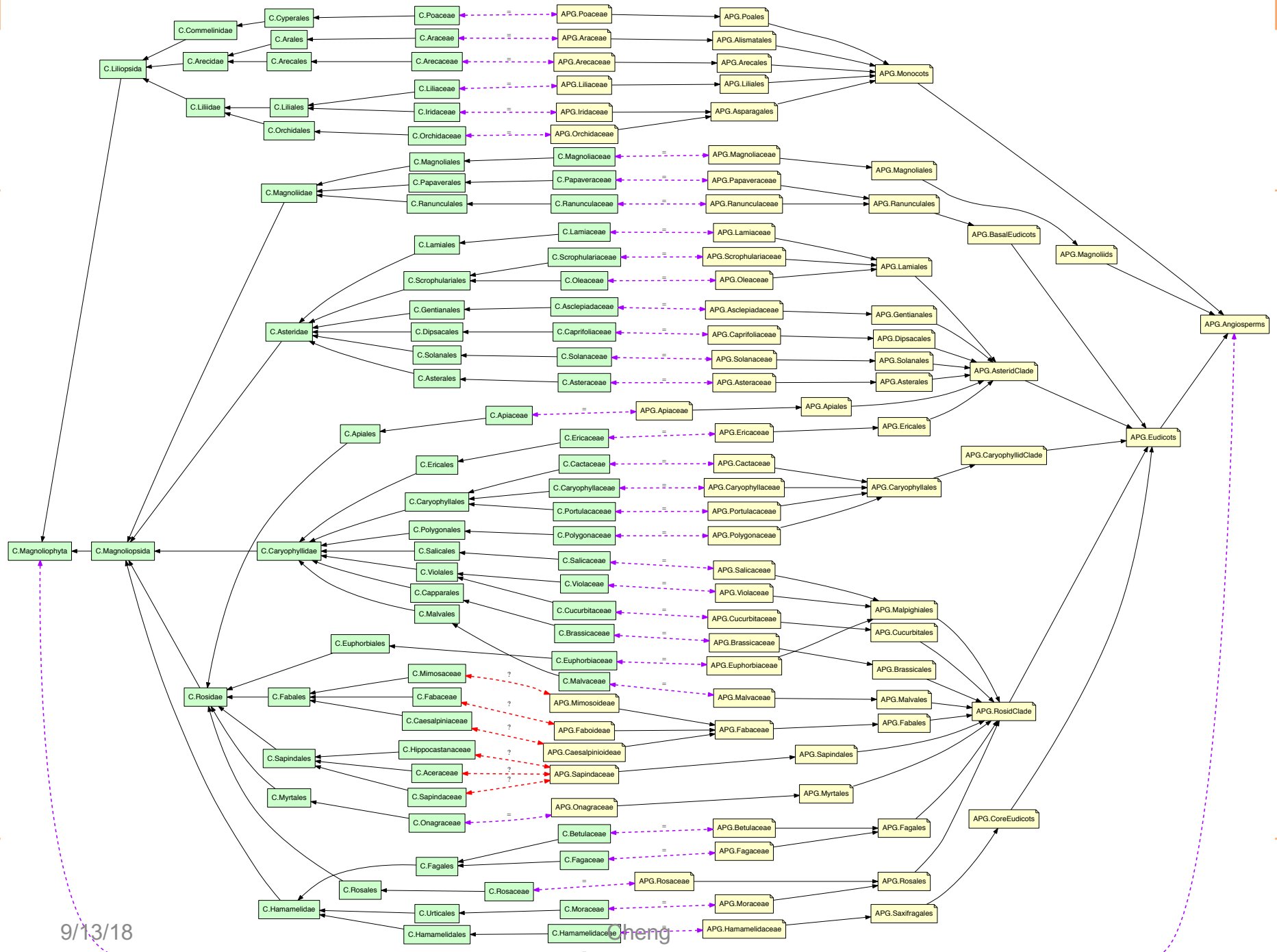
Zooming in to the input alignments
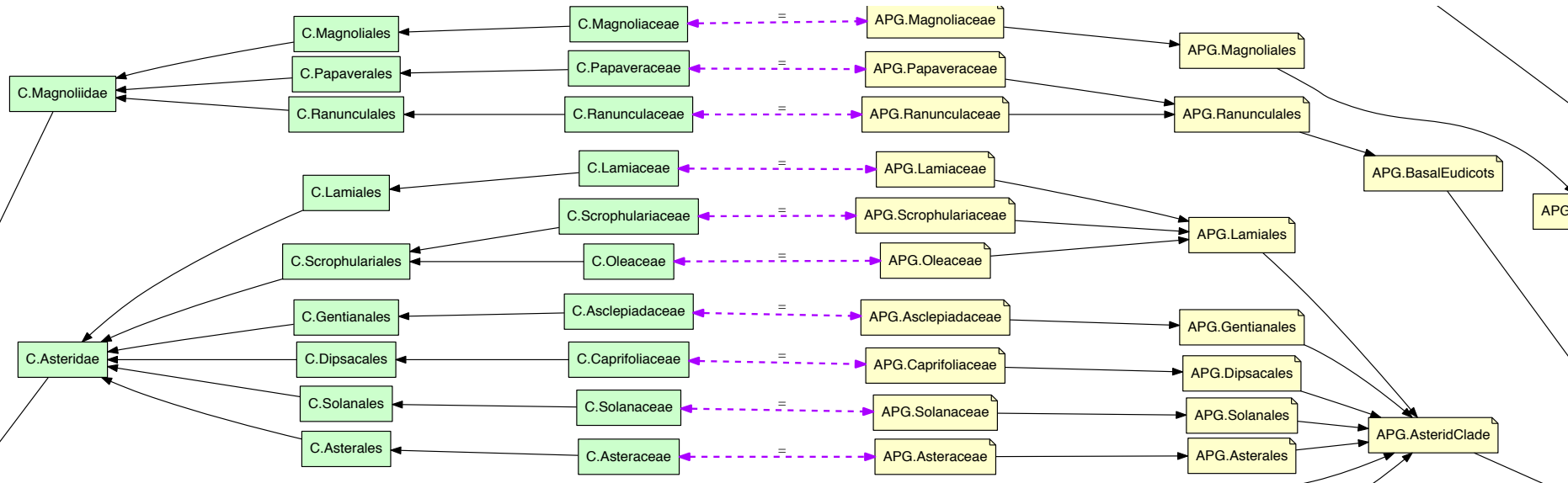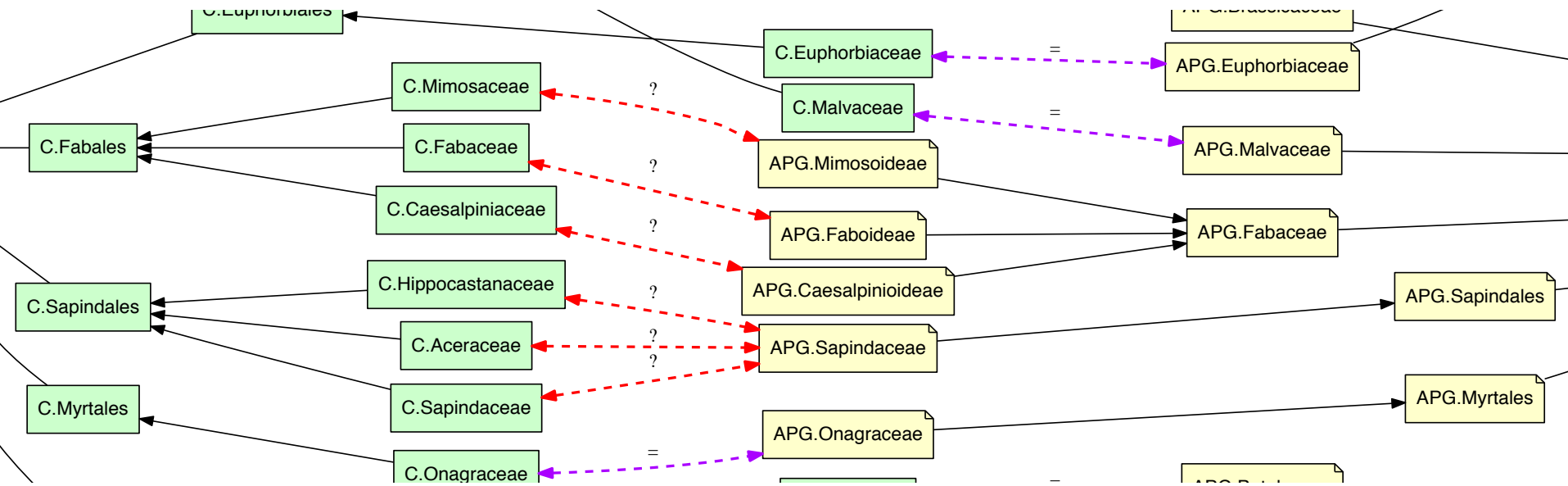
Cheng

School of
Information Sciences
The iSchool at Illinois

Six unknown relations

[C.Caesalpin**iaceae** ? APG.Caesalpin**ioideae**]
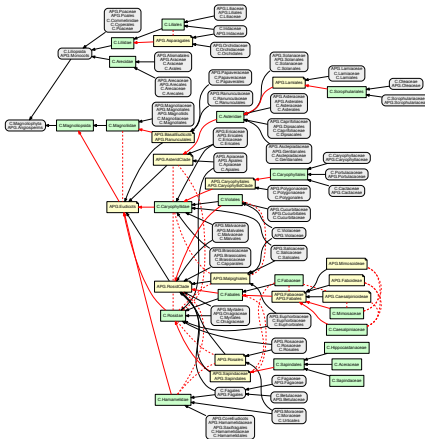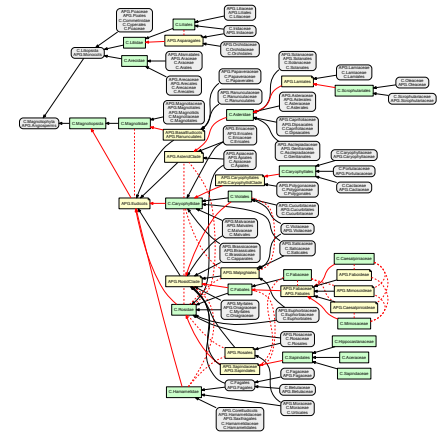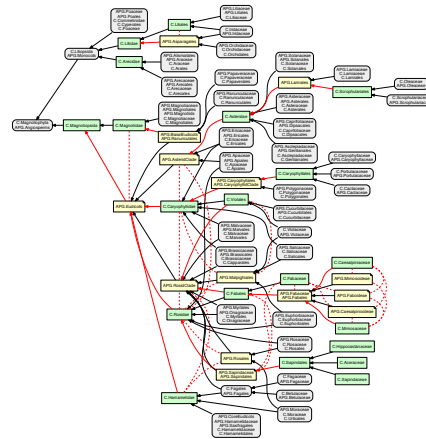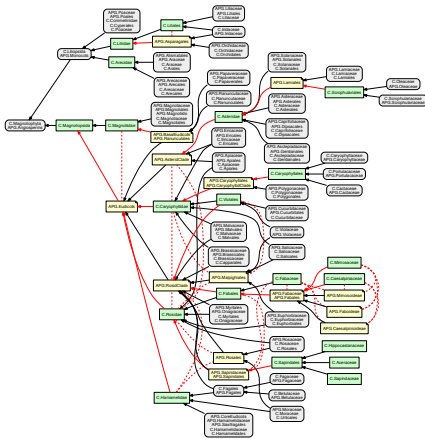[C.Mimosa**ceae** ? APG.Mimosoi**deae**]
[C.Faba**ceae** ? APG.Faboi**deae**]
[C.Aceraceae ? APG.**Sapindaceae**]
[C.Sapindaceae ? APG.**Sapindaceae**]
[C.Hippocastanaceae ? APG.**Sapindaceae**]

Cheng

School of
Information Sciences
The iSchool at Illinois

Results: 555 Possible Worlds

School of
Information Sciences
The iSchool at Illinois

Zooming in..
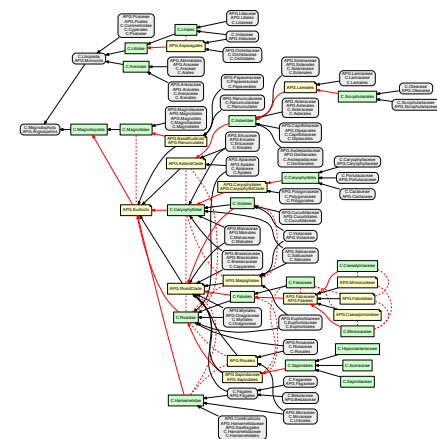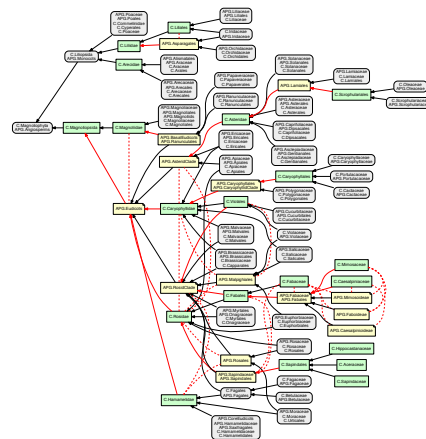
Second round of alignment

- Expert's input
  - To verify our 'congruent' alignments as well as to sort out the underspecified articulations

- Minimum viable product
  - The results from the first stage of alignment help us communicate with the expert and let him/her grasp all possible solutions for the alignment problem

School of
Information Sciences
The iSchool at Illinois

Refining the six unknown relations
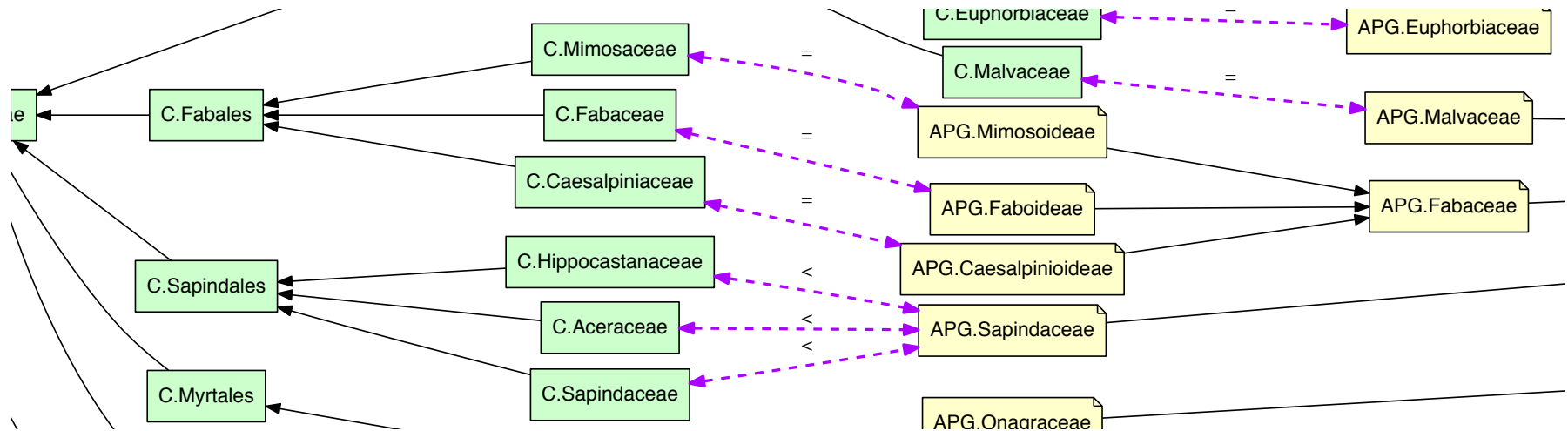
[C.Caesalpiniaceae {=} APG.Caesalpinioideae]

[C.Mimosaceae {=} APG.Mimosoideae]

[C.Fabaceae {=} APG.Faboideae]

[C.Aceraceae {<} APG.Sapindaceae]

[C.Sapindaceae {<} APG.Sapindaceae]

[C.Hippocastanaceae {<} APG.Sapindaceae]

School of
Information Sciences
The iSchool at Illinois

Final result: One Possible World

Cheng

School of
Information Sciences
The iSchool at Illinois
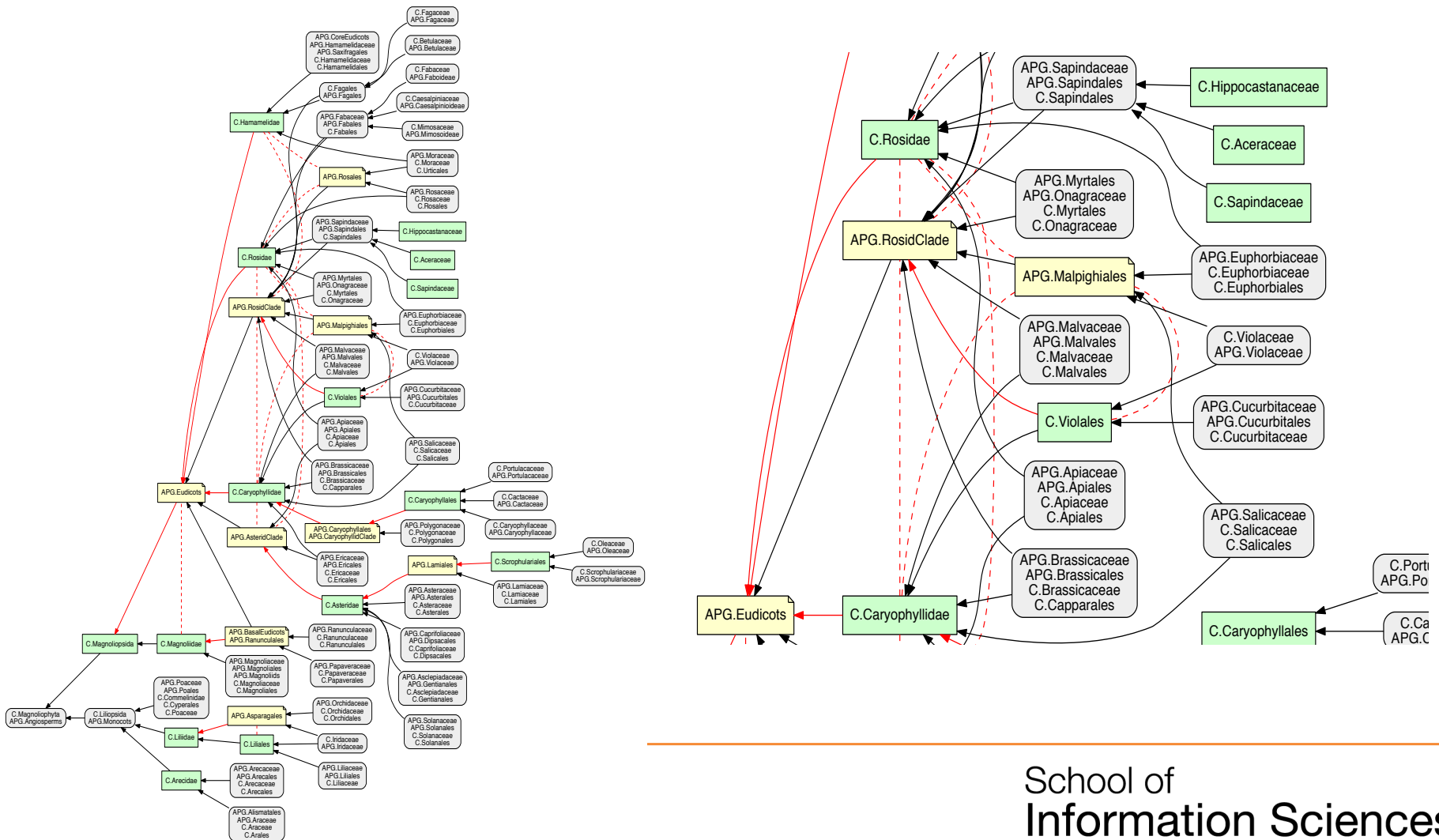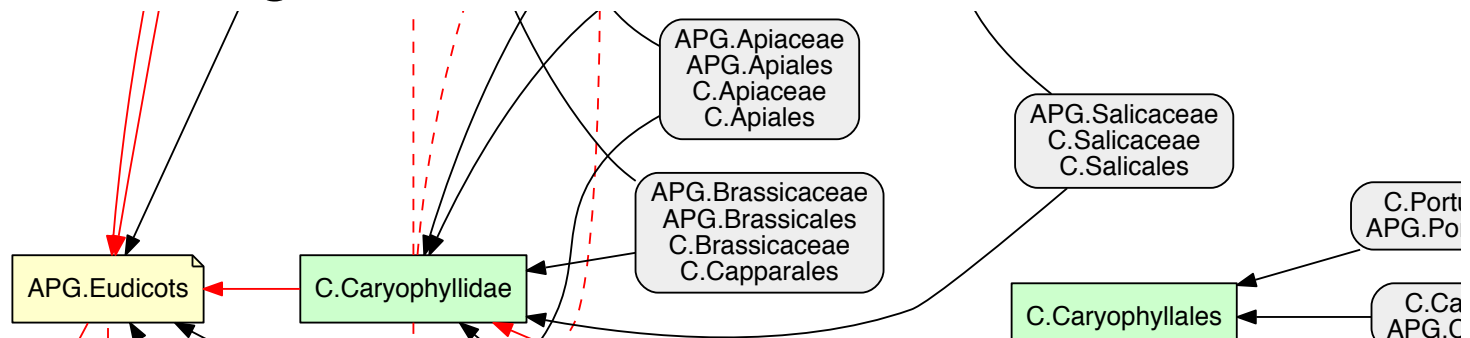
Limitations

- When a parent node only has one child, the RCC reasoner in Euler/X will collapse the concepts and merge them as the same node

- Our response: we have only chosen some 40 major flower families instead of all 416 families. We could add missing children or artificial children here

Cheng

School of
Information Sciences
The iSchool at Illinois

Conclusion and discussion

- Why using the logic-based RCC-5, Euler/X approach?
  - classifications can coexist & disambiguating the names among concepts in a merged possible world
  - solving complex alignments for cases where manual efforts would likely fail to yield all 555 different ways to merge and reconcile the two KOS
- Domain experts seems very important..?
  - KOS alignment problems are usually complex-- our logic-based alignment can serve as ***minimum viable product***

School of
Information Sciences
The iSchool at Illinois

Conclusion and discussion

- Why bother aligning the 'older' classifications such as the Cronquist system (1981)?
  - the Cronquist system still maintains its esteemed role for its comprehensiveness and preciseness in morphologically classifying the flowering plants

- Some other implications
  - making our classification systems more "full of beans"
  - to enable semantic interoperability, and enrich diversity in classification systems

Cheng

School of
Information Sciences
The iSchool at Illinois

Thank you!

Github Repo:
https://github.com/yiyunyc2/NKOS18

Yi-Yun (Jessica) Cheng
yiyunyc2@illinois.edu
Twitter: @yiyunjessica