String Similarity Small Group Outcome

ICANN

Agenda

- Overview
- Task 1
- Task 2
- Task 3
- Conclusion

Overview

Background

Charter Questions

EPDP-IDN Charter asks to consider any adjustment to the string similarity review due to the variant implementation: (Charter Question E3)

- What role, if any, do the “withheld same entity” variants play? (Charter Question E1)
- What are the potential consequences for the other allocatable variant labels in the same set of a requested variant label, which is rejected as a result of the string similarity review? (Charter Question E3a)

Staff Paper Recommendation

String similarity review should compare strings under consideration not just against all allocated or applied-for strings, but also all variants of those strings (i.e., allocatable, withheld-same-entity, and blocked).

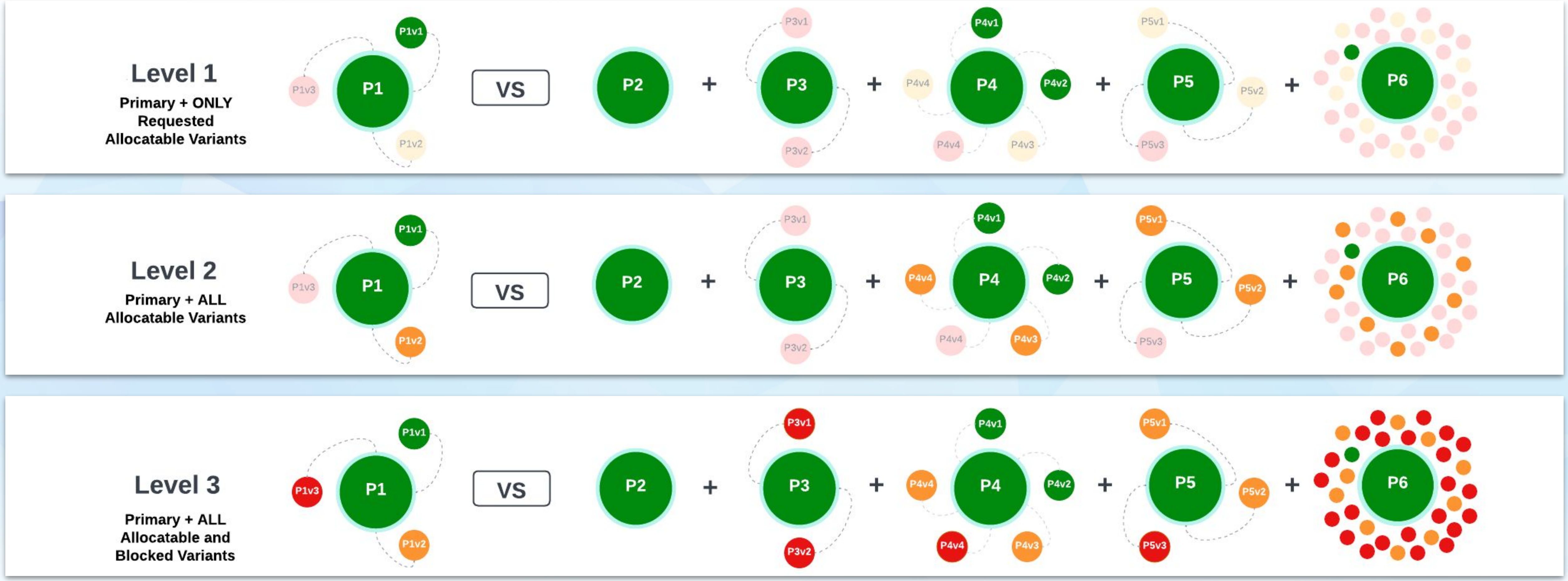
EPDP Team Discussion

The EPDP Team discussed three (3) possible levels of comparison among visually confusable strings, as well as analyzed the impact and potential consequences:

Level 1: Primary + only requested allocatable variants
Level 2: Primary + all allocatable variants
Level 3: Source gTLD + all valid variants (blocked + allocatable)

***Primary**: The applied-for gTLD that also serves as the source gTLD for calculating its allocatable and blocked variants during the application process; the applicant may request to activate none, one, or more allocatable variants of such an applied-for gTLD.*

Three Levels of Comparison



Level 1

Primary + ONLY Requested Allocatable Variants

Level 2

Primary + ALL Allocatable Variants

Level 3

Primary + ALL Allocatable and Blocked Variants

🟢 Requested Allocatable Label

🟠 Non-Requested Allocatable Label

🔴 Blocked Label

Problem Statement

String Similarity Small Group has been set up to tackle the following problems:

Problem 1: There is a divergence of opinions regarding which level is the most appropriate

Problem 2: The discussion has been largely academic based on abstract concepts

Small Group Tasks

Facilitate a comprehensible discussion by ***developing concrete examples of variants that are visually confusable***

Task 1: Develop ***concrete examples of strings*** that have blocked and/or allocatable variant labels and may be visually confusable with other strings in the same scripts or across scripts

- Develop practical examples – limit to visual similarity – that could happen in reality & indicate how feasible/possible such cases could happen
- Discuss whether any existing mechanisms that could help prevent such confusingly similar strings being delegated

Task 2: Demonstrate ***how these examples would be compared against each other in the string similarity review according to the three levels***, showcasing the impact on the review and the potential consequences

- Propose a String Similarity Review model with the view of minimizing security, stability, and user confusability risks

Task 3: Demonstrate ***how these examples would undergo the objection process according to the three levels***, showcasing the impact on the objection process and the potential consequences

- Identify which type of variants should be subject to the objection process

Exclusion: Complexity in implementation for Tasks 2 and 3 is out of scope – defer deliberation to EPDP Team

Small Group Composition

Member	Affiliation	Language Proficiency
Edmon Chung	Board Liaison	Chinese (Mandarin, Cantonese)
Hadia El miniawi	ALAC	Arabic
Imran Hossen	Independent	Bangla
Jerry Sen	RySG	Chinese (Mandarin)
Justine Chew (Small Group Lead)	ALAC	Malay
Michael Bauland	RrSG	German
Wael Nasr	Independent	Arabic

Note:

- Between 18 May 2022 and 3 August 2022, the Small Group held a total of 10 meetings
- Small Group agreed to the 3 tasks stated in the assignment form during its first meeting on 18 May 2022
- Supported by ICANN staff with additional language proficiency
- Wael Nasr joined toward the end of small group deliberation

Task 1

Develop concrete examples of strings that have blocked and/or allocatable variant labels and may be visually confusable with other strings in the same scripts or across scripts

Example Strings

The group developed **eight (8) examples**, as contributed by both members and staff, and discussed their **primary**, **allocatable** and, **blocked** variants calculated by RZ-LGR

No.	Label A	Label B	Label C	Practicality Consideration
1	Latin **bıß**	Cyrillic **вiss**		Valid strings per RZ-LGR
2	Traditional Chinese **滙豐**	Simplified Chinese **汇丰**		Real Chinese words with same meanings and variant relationship
3	Arabic بنئ	Arabic بنی		Valid strings per RZ-LGR with at least one string that's meaningful in a language
4	Simplified Chinese **华鸟**	Traditional Chinese **华島**		Real Chinese words with different meanings
5	Latin **rıch**	Latin **ṅch**		Valid strings per RZ-LGR
6	Arabic رکی	Arabic رکے		Valid strings per RZ-LGR with at least one string that's meaningful in a language
7	Simplified Chinese **华为**	Simplified Chinese **华鸟**	Simplified Chinese **华岛**	Real Chinese words with different meanings
8	Japanese Kanji **一休**	Traditional Chinese **一體**		Real Japanese and Chinese words with different meanings

Task 2

Demonstrate how these examples would be compared against each other in the string similarity review according to the three levels, showcasing the impact on the review and the potential consequences

Selected Examples for Comparison

No.	Label A	Label B	Label C
1	Latin **bıß**	Cyrillic **віѕѕ**	
2	Traditional Chinese **滙豐**	Simplified Chinese **汇丰**	
3	Arabic بنئ	Arabic بنی	
4	Simplified Chinese **华鸟**	Traditional Chinese **华島**	
5	Latin **rıch**	Latin **ṅch**	
6	**Arabic رکی**	**Arabic رکے**	
7	**Simplified Chinese 华为**	**Simplified Chinese 华鸟**	**Simplified Chinese 华岛**
8	Japanese Kanji **一休**	Traditional Chinese **一體**	

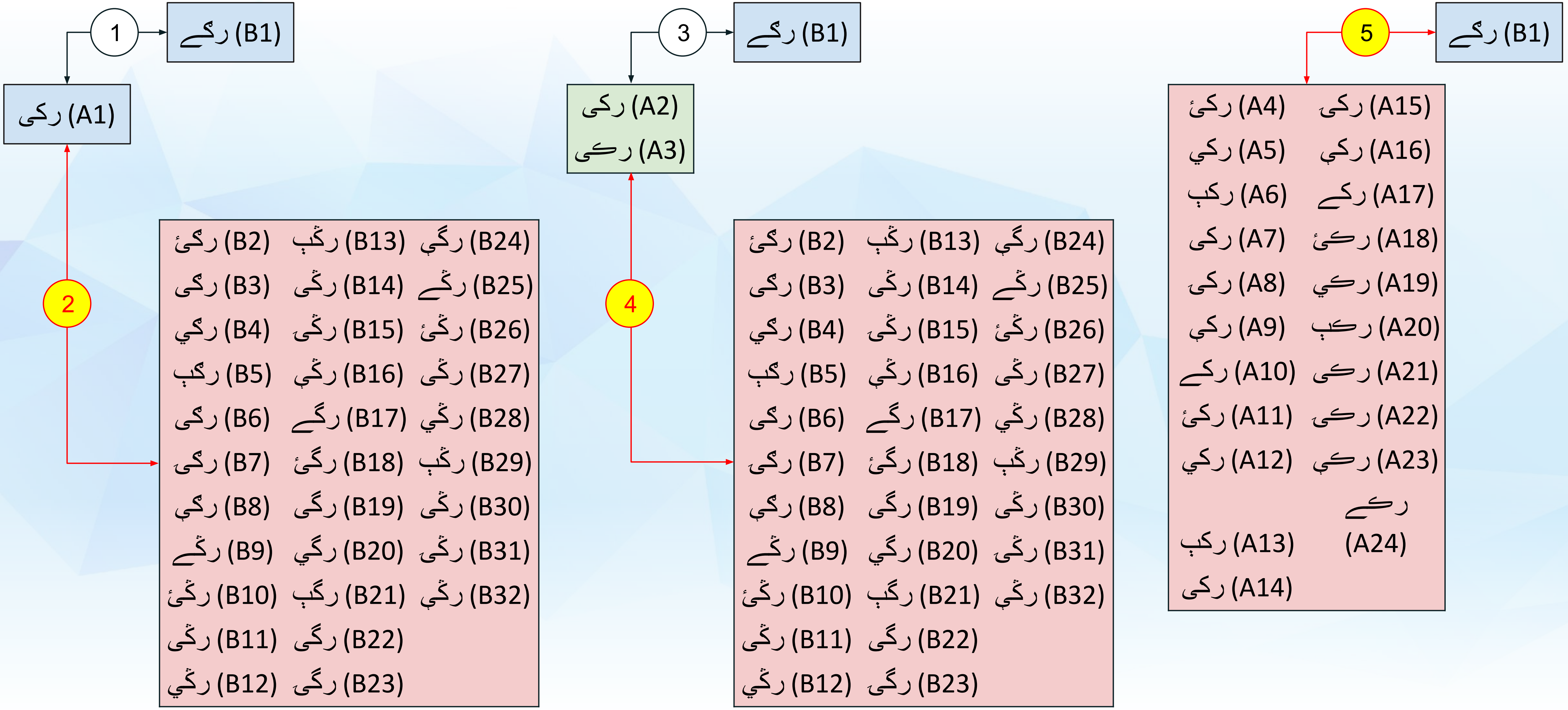
Demonstrate why hybrid model is recommended

Demonstrate
- Applied-for gTLD vs. Existing gTLD
- Comparison among three strings

Example 6: Two Applied-for Arabic TLDs

Applied-for Primary Strings: 🟦	رکی (A1)	رګے (B1)
Allocatable Variants of Primary Strings: 🟩	ركی (A2) رڪی (A3)	None
Blocked Variants of Primary Strings: 🟥	ركئ (A4) ركي (A5) ركې‍ (A6) ركى (A7) ركۍ (A8) ركې (A9) ركے (A10) رکئ (A11) رکي (A12) رکې‍ (A13) رکى (A14) رکۍ (A15) رکې (A16) رکے (A17) رڪئ (A18) رڪي (A19) رڪې‍ (A20) رڪى (A21) رڪۍ (A22) رڪې (A23) رڪے (A24)	رګئ (B2) رګی (B3) رګي (B4) رګې‍ (B5) رګى (B6) رګۍ (B7) رګې (B8) رڭے (B9) رڭئ (B10) رڭی (B11) رڭي (B12) رڭې‍ (B13) رڭى (B14) رڭۍ (B15) رڭې (B16) رگے (B17) رگئ (B18) رگی (B19) رگي (B20) رگې‍ (B21) رگى (B22) رگۍ (B23) رگې (B24) رڴے (B25) رڴئ (B26) رڴی (B27) رڴي (B28) رڴې‍ (B29) رڴى (B30) رڴۍ (B31) رڴې (B32)

Example 6: String Similarity Review



1

رګے (B1)

رکی (A1)

2

رګئ (B2) رڭې‍ (B13) رگې (B24)
رګی (B3) رڭی (B14) رڭے (B25)
رګي (B4) رڭۍ (B15) رڭئ (B26)
رګې‍ (B5) رڭې (B16) رڭی (B27)
رګی (B6) رگے (B17) رڭي (B28)
رګۍ (B7) رگئ (B18) رڭې‍ (B29)
رګې (B8) رگی (B19) رڭی (B30)
رڭے (B9) رگي (B20) رڭۍ (B31)
رڭئ (B10) رگې‍ (B21) رڭې (B32)
رڭی (B11) رگی (B22)
رڭي (B12) رگۍ (B23)

3

رګے (B1)

رکی (A2)
رکـی (A3)

4

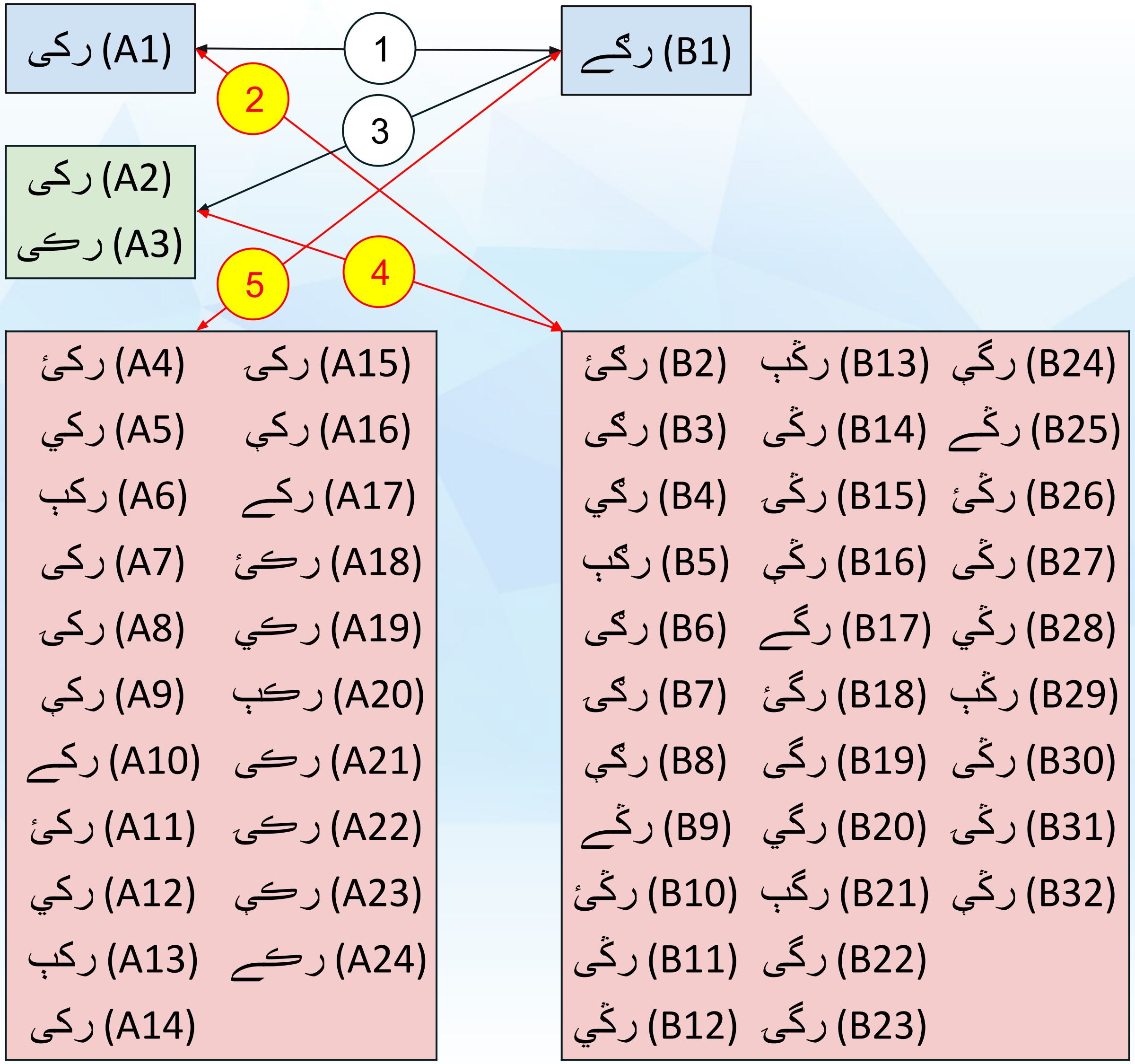
رګئ (B2) رڭې‍ (B13) رگې (B24)
رګی (B3) رڭی (B14) رڭے (B25)
رګي (B4) رڭۍ (B15) رڭئ (B26)
رګې‍ (B5) رڭې (B16) رڭی (B27)
رګی (B6) رگے (B17) رڭي (B28)
رګۍ (B7) رگئ (B18) رڭې‍ (B29)
رګې (B8) رگی (B19) رڭی (B30)
رڭے (B9) رگي (B20) رڭۍ (B31)
رڭئ (B10) رگې‍ (B21) رڭې (B32)
رڭی (B11) رگی (B22)
رڭي (B12) رگۍ (B23)

5

رګے (B1)

رکئ (A4) رکۍ (A15)
رکي (A5) رکې (A16)
رکې‍ (A6) رکے (A17)
رکی (A7) رکـئ (A18)
رکۍ (A8) رکـي (A19)
رکې (A9) رکـې‍ (A20)
رکے (A10) رکـی (A21)
رکئ (A11) رکـۍ (A22)
رکي (A12) رکـې (A23)
رکې‍ (A13) رکـے (A24)
رکی (A14)

Example 6: String Similarity Review (Cont.)



رکئ (A4)
رکي (A5)
رکۑ (A6)
رکى (A7)
رکۍ (A8)
رکې (A9)
رکے (A10)
ركئ (A11)
ركي (A12)
ركۑ (A13)
ركى (A14)
ركۍ (A15)
ركې (A16)
ركے (A17)
رڪئ (A18)
رڪي (A19)
رڪۑ (A20)
رڪى (A21)
رڪۍ (A22)
رڪې (A23)
رڪے (A24)

رګئ (B2)
رګی (B3)
رګي (B4)
رګۑ (B5)
رګى (B6)
رګۍ (B7)
رګې (B8)
رڭے (B9)
رڭئ (B10)
رڭی (B11)
رڭي (B12)
رڭۑ (B13)
رڭى (B14)
رڭۍ (B15)
رڭې (B16)
رگے (B17)
رگئ (B18)
رگی (B19)
رگي (B20)
رگۑ (B21)
رگى (B22)
رگۍ (B23)
رگې (B24)
رݣے (B25)
رݣئ (B26)
رݣی (B27)
رݣي (B28)
رݣۑ (B29)
رݣى (B30)
رݣۍ (B31)
رݣې (B32)

String Similarity Review may find the following confusingly similar strings

2 رکی (A1) & رګی (B3) & رګى (B6)

4 ركی (A2) & رګی (B3) & رګى (B6)

4 رڪی (A3) & رګی (B3) & رګى (B6)

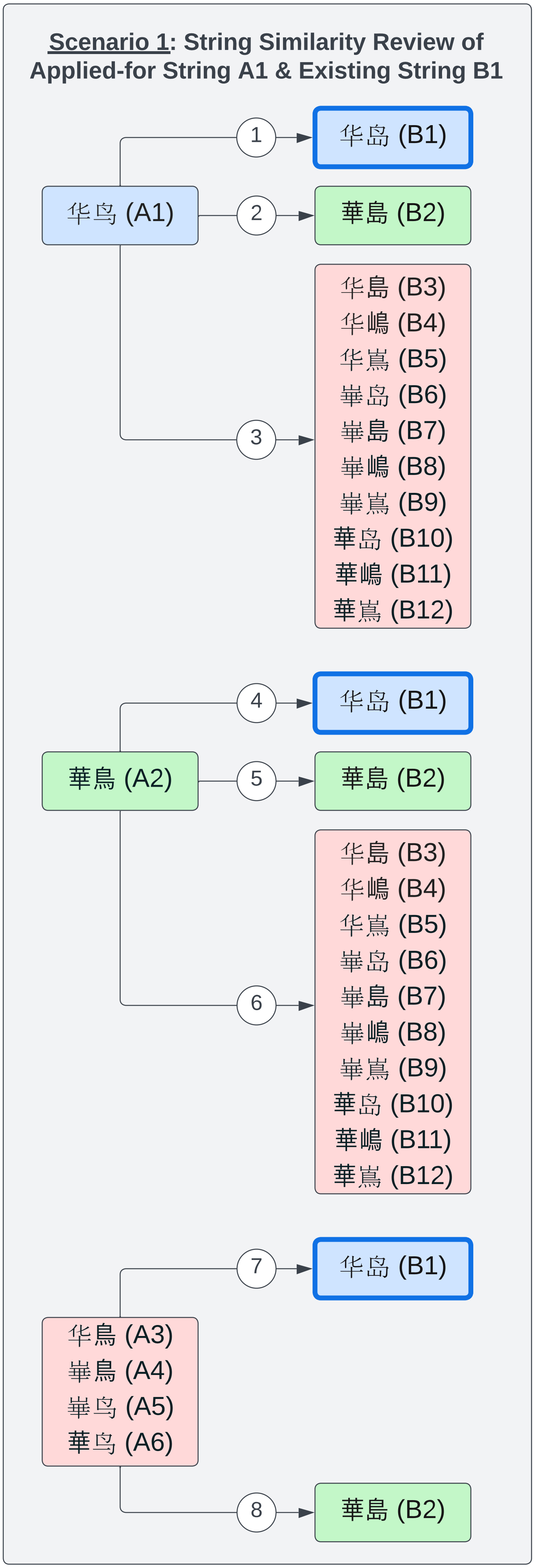
5 رګے (B1) & رکے (A10) & ركے (A17) & رڪے (A24)

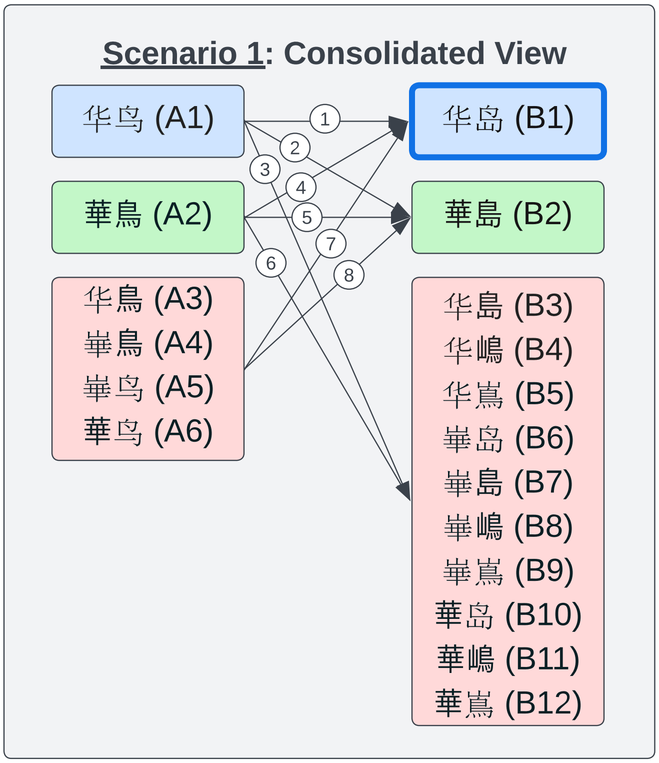
Potential Outcome of the String Similarity Review

رکی (A1) & its variants A2-A24 AND رګے (B1) & its variants B2-B32 get processed in a contention set

If the hybrid model were not used and blocked variants were not taken into account in String Similarity Review

رکی (A1) and رګے (B1) would have been both delegated with the misconnection risk. E.g., a user may mistake رکی (A1) as رګی (B3), a blocked variant of رګے (B1), but arrive at site controlled by a registrant different to رګے (B1).





String Similarity Review may find the following confusingly similar pairs

1 华鸟 (A1) & 华岛 (B1)

5 華鳥 (A2) & 華島 (B2)

6 華鳥 (A2) & 崋島 (B7)

8 崋鳥 (A4) & 華島 (B2)

Potential Outcome of String Similarity Review

华鸟 (A1) may be rejected due to its confusing similarity to the already delegated 华岛 (B1)

Legend

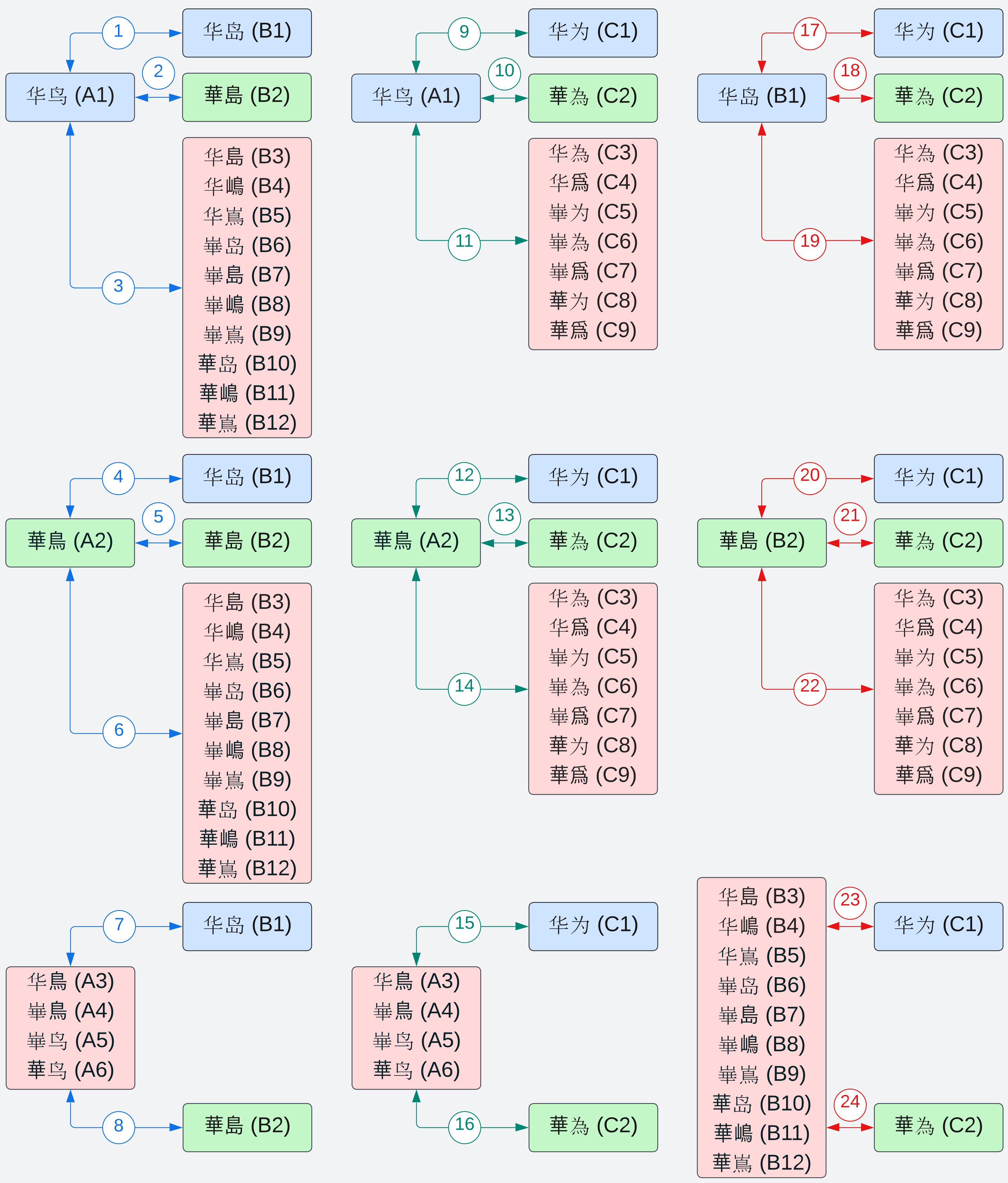
🟦 Primary applied-for string

🟦 Primary existing string

🟩 Allocatable variant of primary string

🟥 Blocked variant of primary string

Scenario 2: String Similarity Review of Applied-for Strings A1, B1 & C1



1 华鸟 (A1) 华岛 (B1)
2 华鸟 (A1) 華島 (B2)
3 华鸟 (A1) 华島 (B3) 华嶋 (B4) 华嶌 (B5) 崋岛 (B6) 崋島 (B7) 崋嶋 (B8) 崋嶌 (B9) 華岛 (B10) 華嶋 (B11) 華嶌 (B12)

9 华鸟 (A1) 华为 (C1)
10 华鸟 (A1) 華為 (C2)
11 华鸟 (A1) 华為 (C3) 华爲 (C4) 崋为 (C5) 崋為 (C6) 崋爲 (C7) 華为 (C8) 華爲 (C9)

17 华岛 (B1) 华为 (C1)
18 华岛 (B1) 華為 (C2)
19 华岛 (B1) 华為 (C3) 华爲 (C4) 崋为 (C5) 崋為 (C6) 崋爲 (C7) 華为 (C8) 華爲 (C9)

4 華鳥 (A2) 华岛 (B1)
5 華鳥 (A2) 華島 (B2)
6 華鳥 (A2) 华島 (B3) 华嶋 (B4) 华嶌 (B5) 崋岛 (B6) 崋島 (B7) 崋嶋 (B8) 崋嶌 (B9) 華岛 (B10) 華嶋 (B11) 華嶌 (B12)

12 華鳥 (A2) 华为 (C1)
13 華鳥 (A2) 華為 (C2)
14 華鳥 (A2) 华為 (C3) 华爲 (C4) 崋为 (C5) 崋為 (C6) 崋爲 (C7) 華为 (C8) 華爲 (C9)

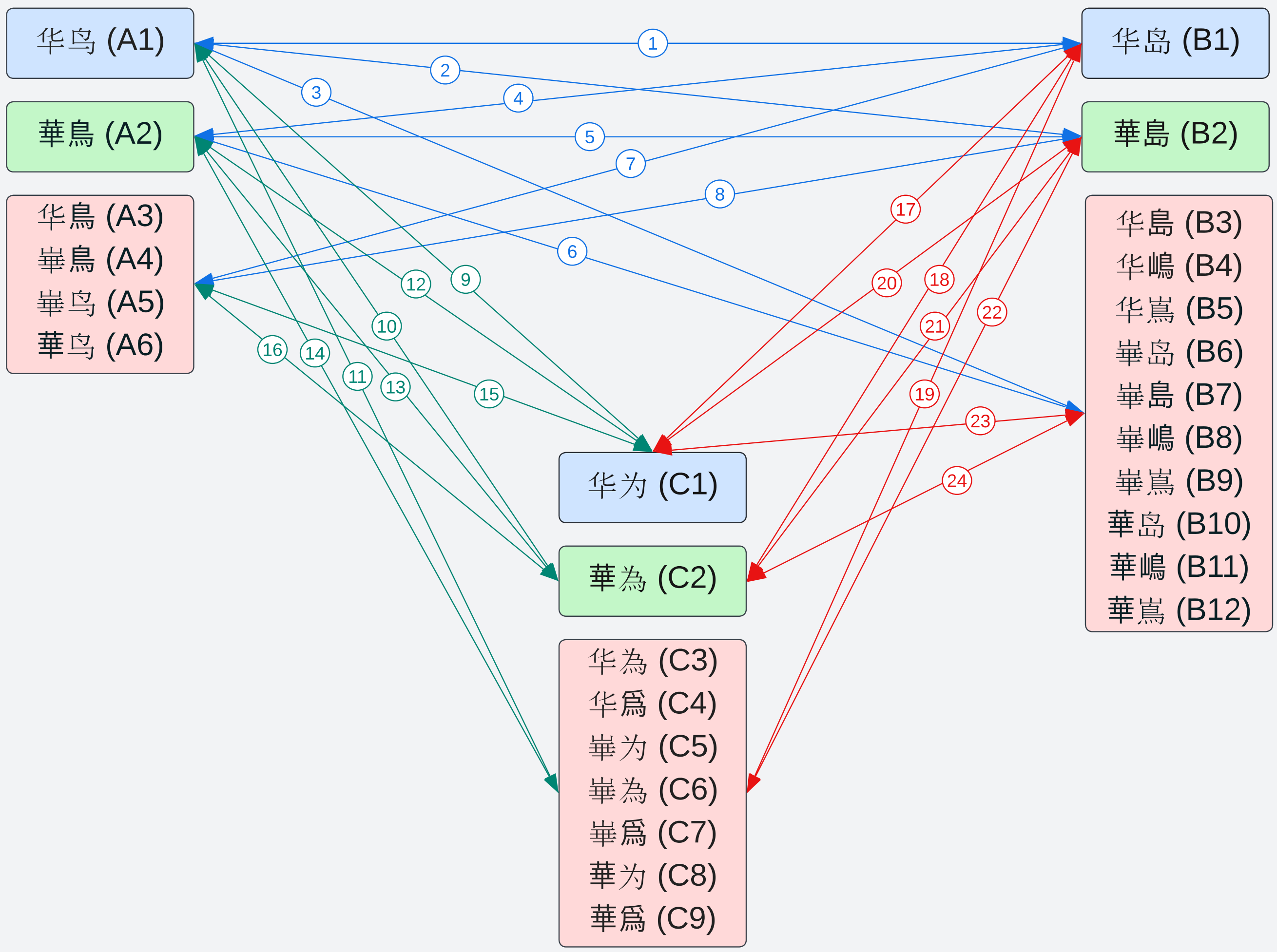
20 華島 (B2) 华为 (C1)
21 華島 (B2) 華為 (C2)
22 華島 (B2) 华為 (C3) 华爲 (C4) 崋为 (C5) 崋為 (C6) 崋爲 (C7) 華为 (C8) 華爲 (C9)

7 华鳥 (A3) 崋鳥 (A4) 崋鸟 (A5) 華鸟 (A6) 华岛 (B1)
8 华鳥 (A3) 崋鳥 (A4) 崋鸟 (A5) 華鸟 (A6) 華島 (B2)

15 华鳥 (A3) 崋鳥 (A4) 崋鸟 (A5) 華鸟 (A6) 华为 (C1)
16 华鳥 (A3) 崋鳥 (A4) 崋鸟 (A5) 華鸟 (A6) 華為 (C2)

23 华島 (B3) 华嶋 (B4) 华嶌 (B5) 崋岛 (B6) 崋島 (B7) 崋嶋 (B8) 崋嶌 (B9) 華岛 (B10) 華嶋 (B11) 華嶌 (B12) 华为 (C1)
24 华島 (B3) 华嶋 (B4) 华嶌 (B5) 崋岛 (B6) 崋島 (B7) 崋嶋 (B8) 崋嶌 (B9) 華岛 (B10) 華嶋 (B11) 華嶌 (B12) 華為 (C2)

Scenario 2: Consolidated View



String Similarity Review may find the following confusingly similar pairs

1 华鸟 (A1) & 华岛 (B1)

5 華鳥 (A2) & 華島 (B2)

6 華鳥 (A2) & 崋島 (B7)

8 崋鳥 (A4) & 華島 (B2)

13 華鳥 (A2) & 華為 (C2)

14 華鳥 (A2) & 崋為 (C6)

16 崋鳥 (A4) & 華為 (C2)

21 華島 (B2) & 華為 (C2)

24 崋島 (B7) & 華為 (C2)

Potential Outcome of String Similarity Review

华鸟 (A1) & its variants A2-A6 AND 华岛 (B1) & its variants B2-B12 AND 华为 (C1) & its variants C2-C9 get processed in a contention set

Legend

🟦 Primary applied-for string

🟩 Allocatable variant of primary string

🟥 Blocked variant of primary string

Recommendation: Hybrid Model

Summary: *The small group recommends the **hybrid model**, which is a **mixed-level approach between level 2 and level 3***

Goal: *Mitigate any possibility of confusing similarity between one IDN TLD and another IDN TLD or any of its valid variant(s), vice versa*

In practice, the string similarity review must be modified to compare:

- **An applied-for primary IDN gTLD and all of its allocatable variant label(s)**

Against:

- **Existing TLDs and all of their allocatable and blocked variant labels;**
- **Strings requested as IDN ccTLDs and all of their allocatable and blocked variant labels;**
- **Other applied-for gTLDs in the same round and all of their allocatable and blocked variant labels;**
- **Reserved Names; and**
- **Any other two-character ASCII strings and all of their allocatable and blocked variant labels (*if the applied-for primary IDN gTLD is a two-character string*)**

In addition, compare:

- **All of the blocked variant label(s) of an applied-for primary IDN gTLD**

Against:

- **Existing TLDs and all of their allocatable variant labels**

Note: *Blocked variants of one IDN TLD should NOT be compared against blocked variants of another IDN TLD*

Rationale for Hybrid Model

Considering the limited scope of security, stability and user confusability, the small group believes the hybrid model would:

- **Be sufficiently conservative** and can **help mitigate two types of failure modes** – denial of service and misconnection, which may have a higher likelihood to affect non-native speakers of certain scripts or languages
- **Help detect many more pairs of visually confusable strings** and **reduce the risks of failure modes**
- **Reduce computational complexity by not requiring comparison of blocked variant labels** of a primary applied-for IDN gTLD string against blocked variant labels of other existing and applied-for TLD strings

The small group also believes that:

- Level 1 and 2 may fail to detect some visually confusable strings and increase the risks of failure modes
- Level 3 unnecessarily compares blocked variants against each other with exponential increase of computational complexity

Additional Considerations

- While the pool of strings that needs to be considered will be large, **language experts in the String Similarity Review panel can evaluate the strings on a case-by-case basis**
- After the evaluation completes, there are **other mechanisms in the New gTLD Program** – e.g., limited appeal mechanism and objection processes – to review the string similarity panel's decision

Conclusion

String Similarity Review Recommendation

Summary:** The small group recommends the **hybrid model**, a **mixed-level approach between level 2 and level 3

The string similarity review must be modified to compare:

- **An applied-for primary IDN gTLD and all of its allocatable variant label(s)**

Against:

- **Existing TLDs and all of their allocatable and blocked variant labels;**
- **Strings requested as IDN ccTLDs and all of their allocatable and blocked variant labels;**
- **Other applied-for gTLDs in the same round and all of their allocatable and blocked variant labels;**
- **Reserved Names; and**
- **Any other two-character ASCII strings and all of their allocatable and blocked variant labels (*if the applied-for primary IDN gTLD is a two-character string*)**

In addition, the string similarity review must be modified to compare:

- **All of the blocked variant label(s) of an applied-for primary IDN gTLD**

Against:

- **Existing TLDs and all of their allocatable variant labels**