

Improving Speaker Identification using Network Knowledge in Criminal Conversational Data

Paper overview

Improving Speaker Identification using Network Knowledge in Criminal Conversational Data

Maël Fabien^{1,2}, Seyyed Saeed Sarfjoo¹, Petr Motlicek¹, Srikanth Madikeri¹

¹Idiap Research Institute, Martigny, Switzerland,

²Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

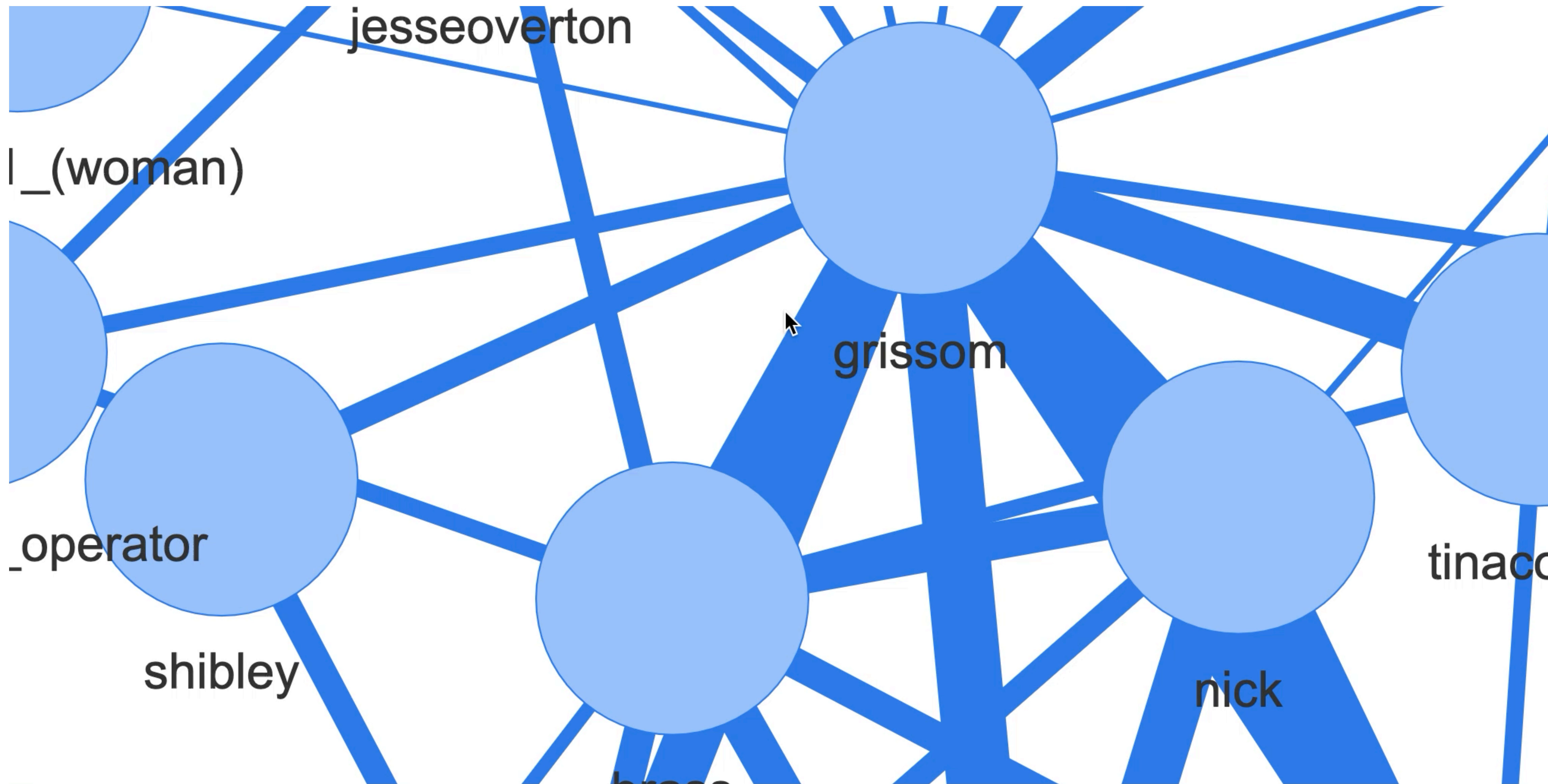
{mfabien, ssarfjoo, petr.motlicek, msrikanth}@idiap.ch

<https://arxiv.org/abs/2006.02093>

Table of Content

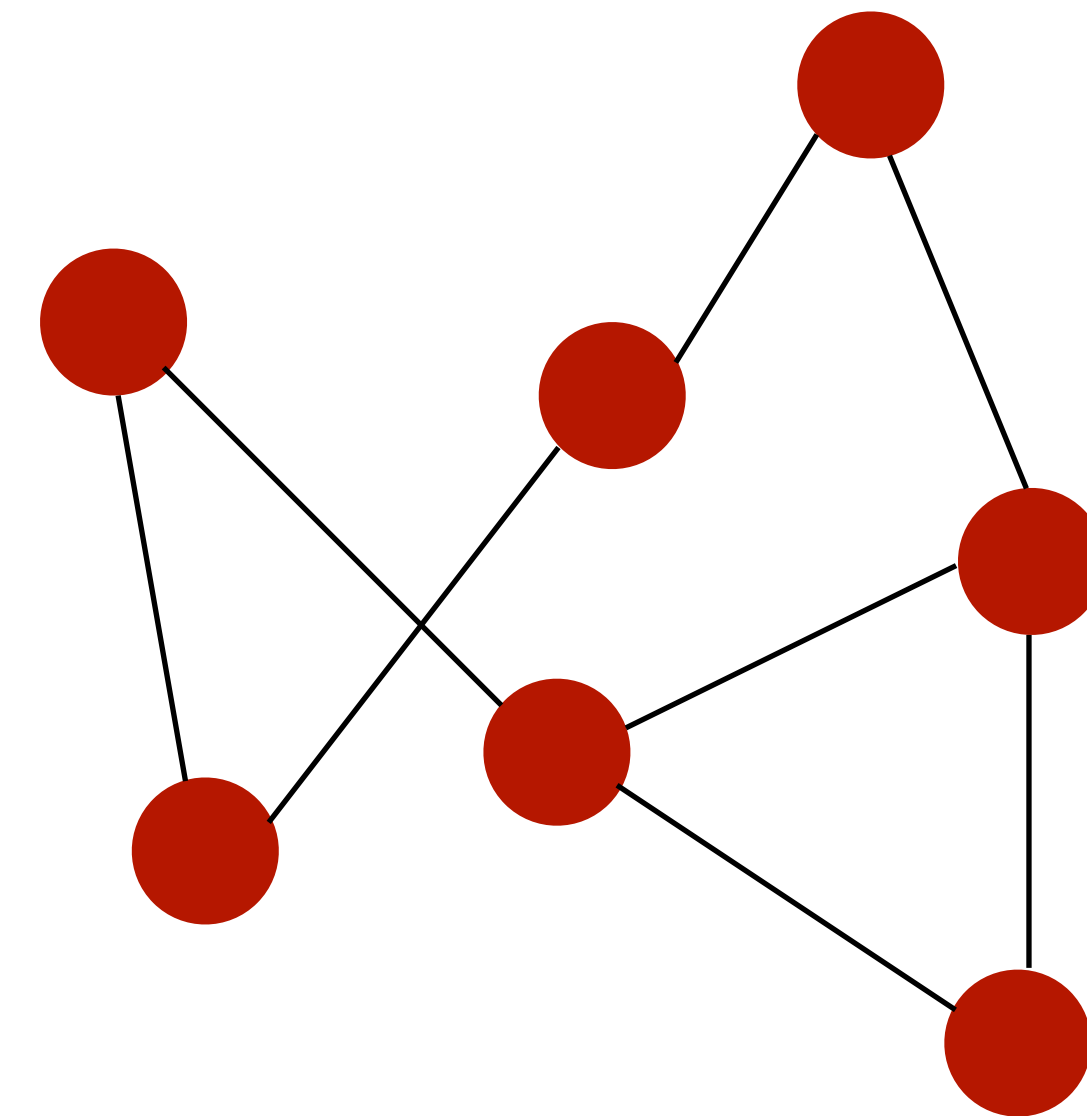
1. Criminal networks and CSI
2. Speaker identification in criminal investigations
3. Evaluation metrics
4. Speaker identification baseline
5. Re-ranking algorithm
6. Results
7. Future works

I. Criminal Networks and CSI



I. Criminal Networks and CSI

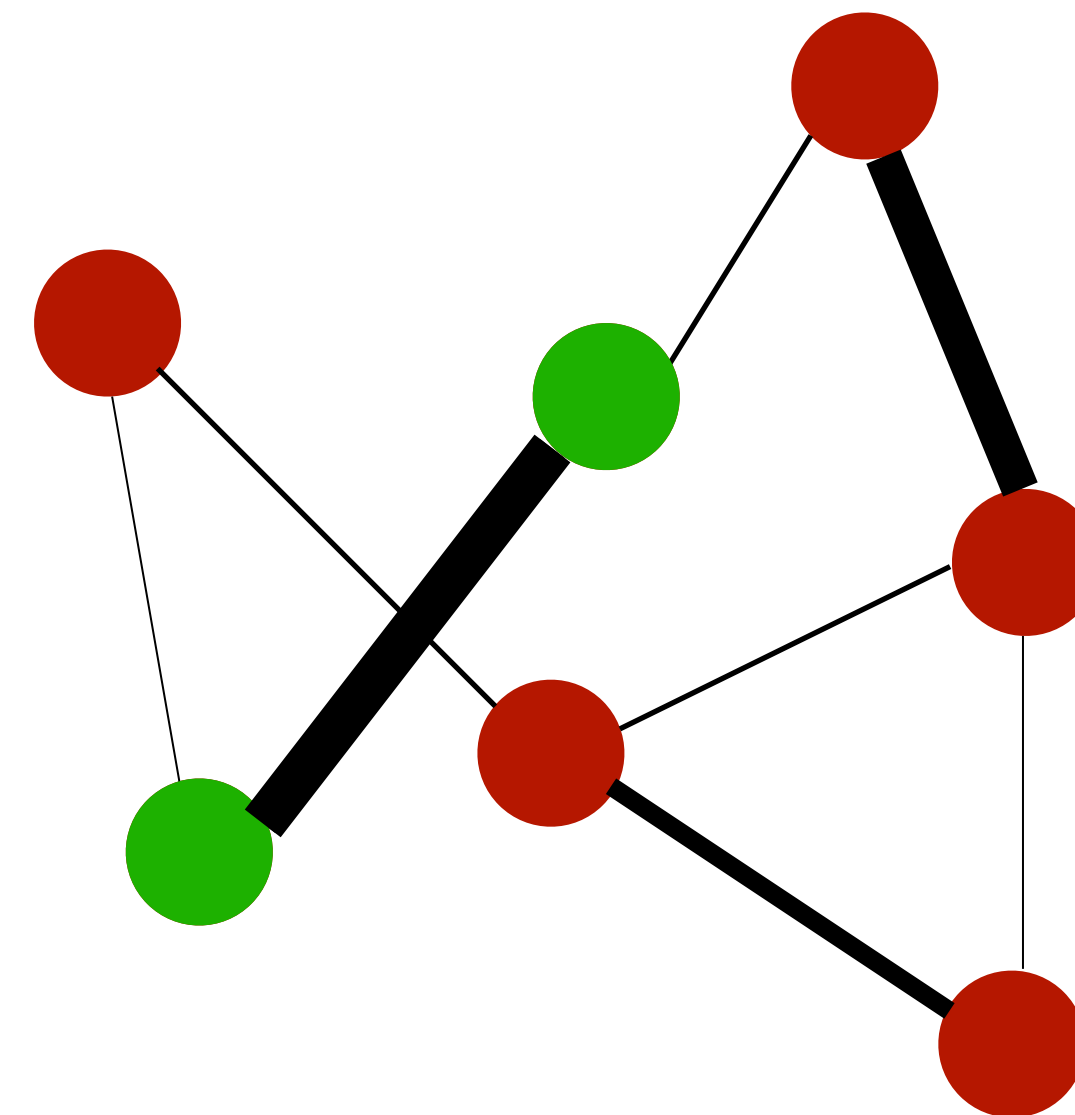
- In real-time investigation condition, as a new audio file is collected, we must assess the identity of the characters (say 2 to keep it simple)
- If we miss-classify one speaker or another, we add a wrong edge in the graph, which can lead investigators on a wrong track



I. Criminal Networks and CSI

- But since we have weighted graphs reflecting previous interactions, we know which link is more likely to be correct

Hypothesis: Criminal networks convey information that could allow us to improve speaker identification in criminal investigations.



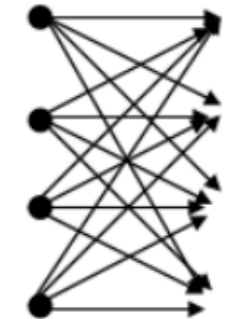
I. Criminal Networks and CSI

e.g. If A and B talk often, and A and B are potential candidates for the identity of potential characters in a conversation, we can re-rank the pairs of characters based on the frequency of previous interactions.

Conversation 4

 Recording 1

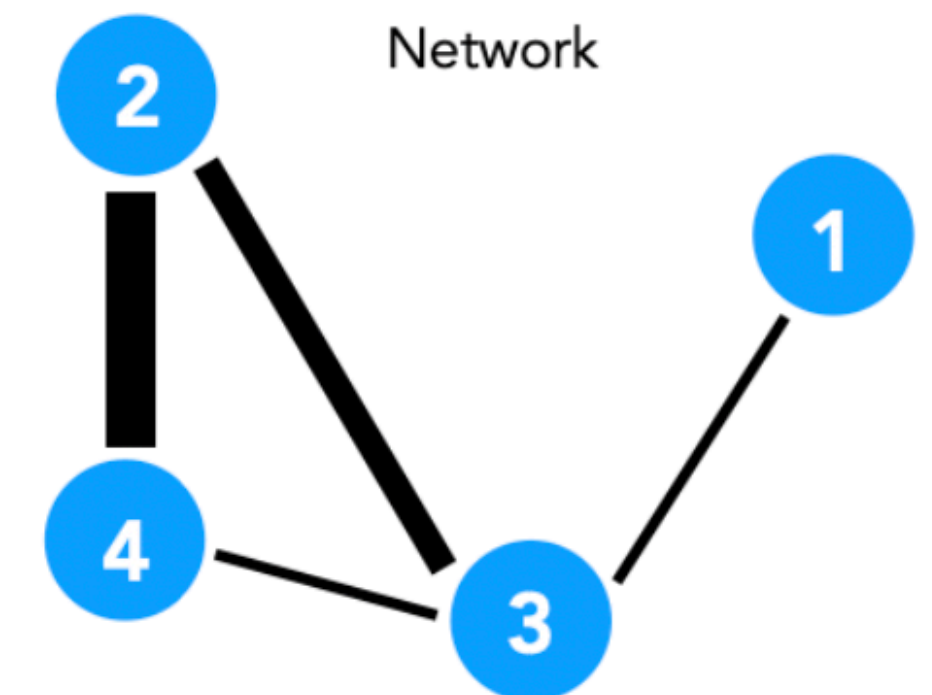
 Recording 2

Speaker_1	14.134		Speaker_1	0.91
Speaker_2	11.032		Speaker_2	1.018
Speaker_3	-11.34		Speaker_3	-11.34
Speaker_4	0.123		Speaker_4	28.19

Speaker 2 and Speaker 4 know each other well, so their score will be higher after re-ranking.

Re-ranking:

Speaker_2 Speaker_4
Speaker_1 Speaker_4
...



I. Criminal Networks and CSI

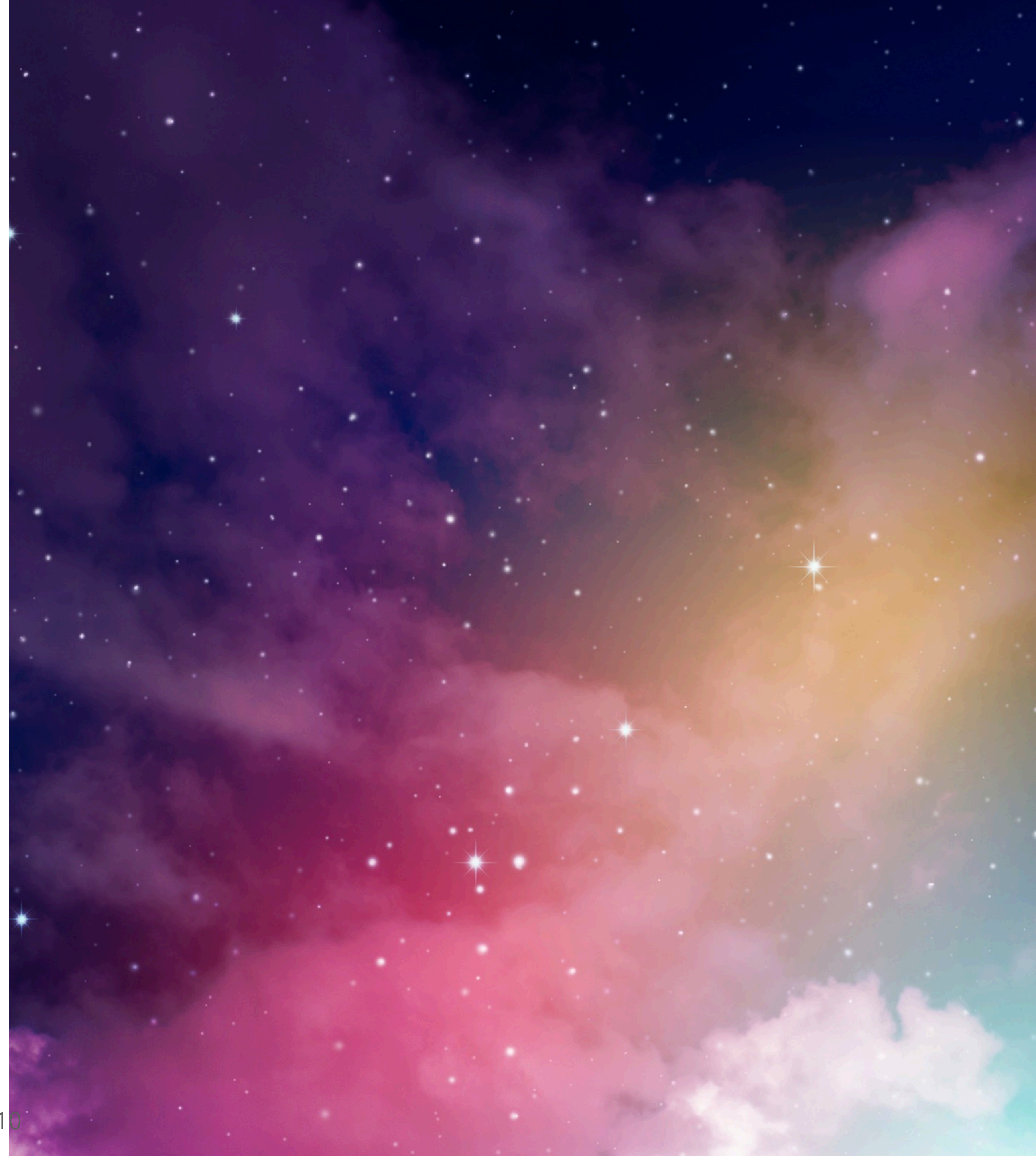
We use Criminal Scene Investigation (CSI) TV-series data:

- transcripts are provided by the University of Edinburgh (<https://github.com/EdinburghNLP/csi-corpus>)
- videos and audio of episodes were extracted from the DVDs we bought

We make the assumption that the topology of the episodes studied in CSI is close enough to criminal investigations.

II.

Speaker identification in criminal investigations



II. Speaker identification in criminal investigations

Existing works: « Leveraging side information for speaker identification with the Enron conversational telephone speech collection. », Ning Gao, Gregory Sell, Douglas W. Oard, Mark Dredze

Steps:

- Speaker diarization using i-vector segments
- Speaker identification using i-vector baseline
- Re-rank the potential speakers in the conversation using past information

$$s_p = \frac{1}{2} \left(\left(1 + \frac{e_l}{\sum e} \right) s_l + \left(1 + \frac{e_r}{\sum e} \right) s_r \right) \left(1 + \frac{e_{lr}}{\sum e} \right)$$

sum of the edge weights connected to the left speaker over all weights

sum of the edge weights connected to the right speaker over all weights

sum of the edge weights between left and right speakers over all weights

acoustic score of left speaker

acoustic score of right speaker

II. Speaker identification in criminal investigations

Existing works: « Leveraging side information for speaker identification with the Enron conversational telephone speech collection. », Ning Gao, Gregory Sell, Douglas W. Oard, Mark Dredze

Acoustic Rank	Ranking of Speaker Pairs	Final Re-ranked List
Speaker01	Speaker01 & Speaker03	Speaker01
Speaker02	Speaker04 & Speaker01	Speaker03
Speaker03		Speaker04
Speaker04		Speaker02

II. Speaker identification in criminal investigations

Results:

- DCF metric is deteriorated
- Classification error metric is improved
- R, the harmonic mean of the rank is improved

Single Source	DCF@0.03	Classification Error	R
Baseline	0.67	0.56	0.73
Email Social Network	0.72	0.49	0.70
Phone Social Network	0.74	0.51	0.65

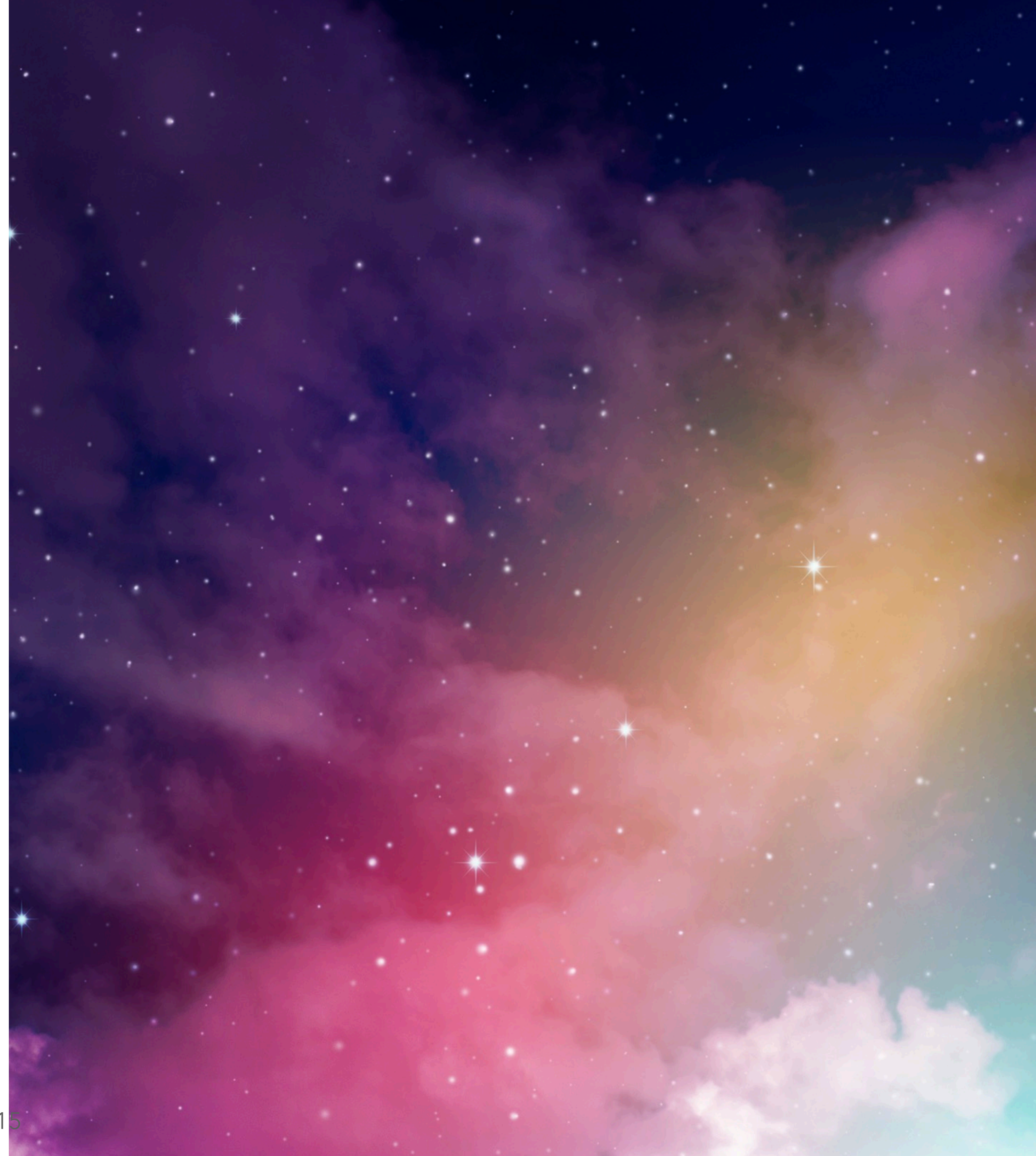
II. Speaker identification in criminal investigations

Limits of the approach:

- Works only if 2 characters are involved in the conversation
- Requires an external source of data (emails) to influence the scores of phone calls

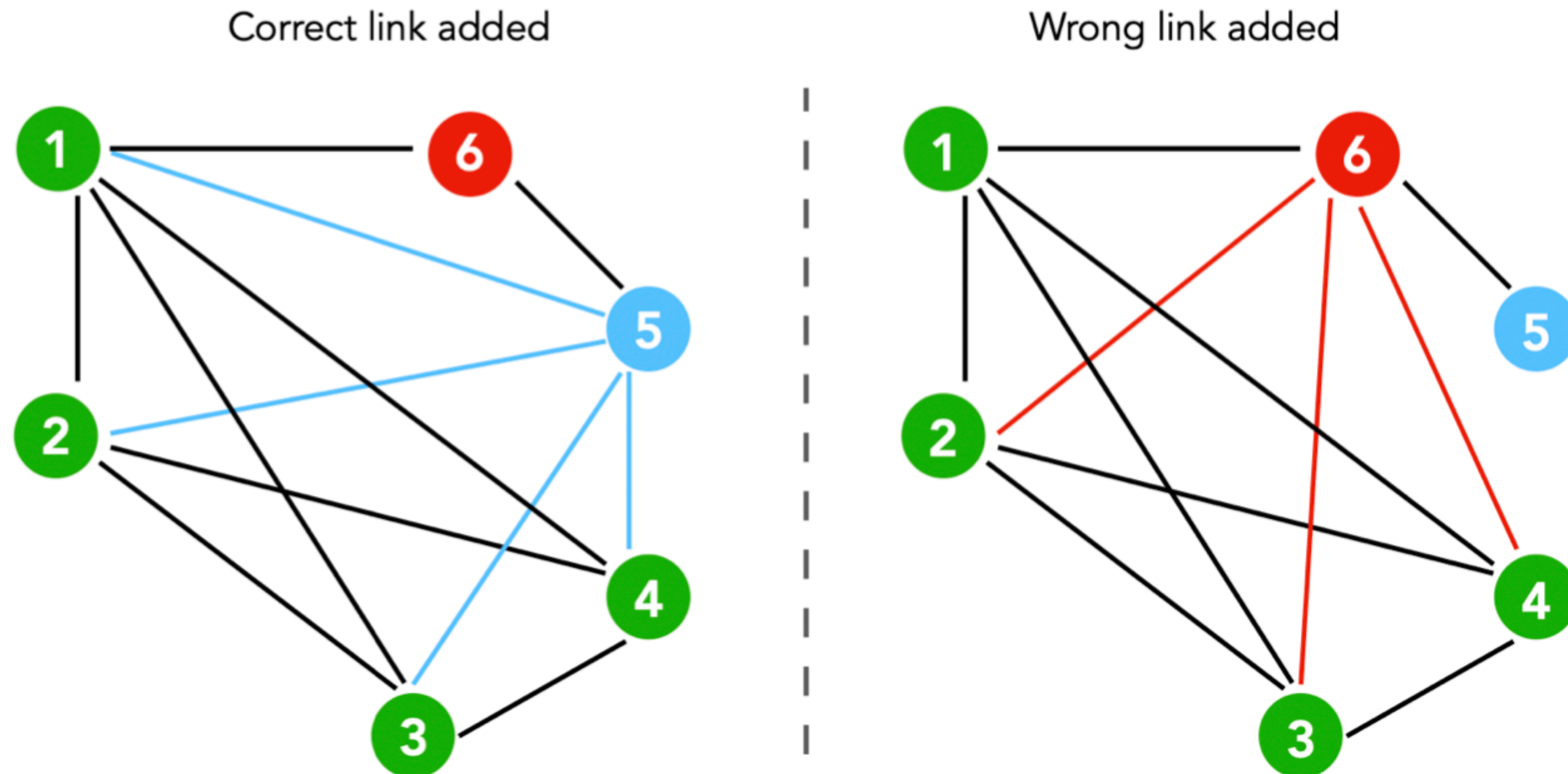
III.

Evaluation metrics



III. Evaluation metrics

Speaker accuracy is a natural metric. However, in a conversation of 5 people, if we mis-identify one character, we add a wrong edge to the network:



III. Evaluation metrics

We introduced the notion of « conversation accuracy », the percentage of conversations for which we correctly identified all characters.

$$acc_C = \frac{1}{C} \sum_{c=1}^C \delta(s_{pc}, s_c),$$

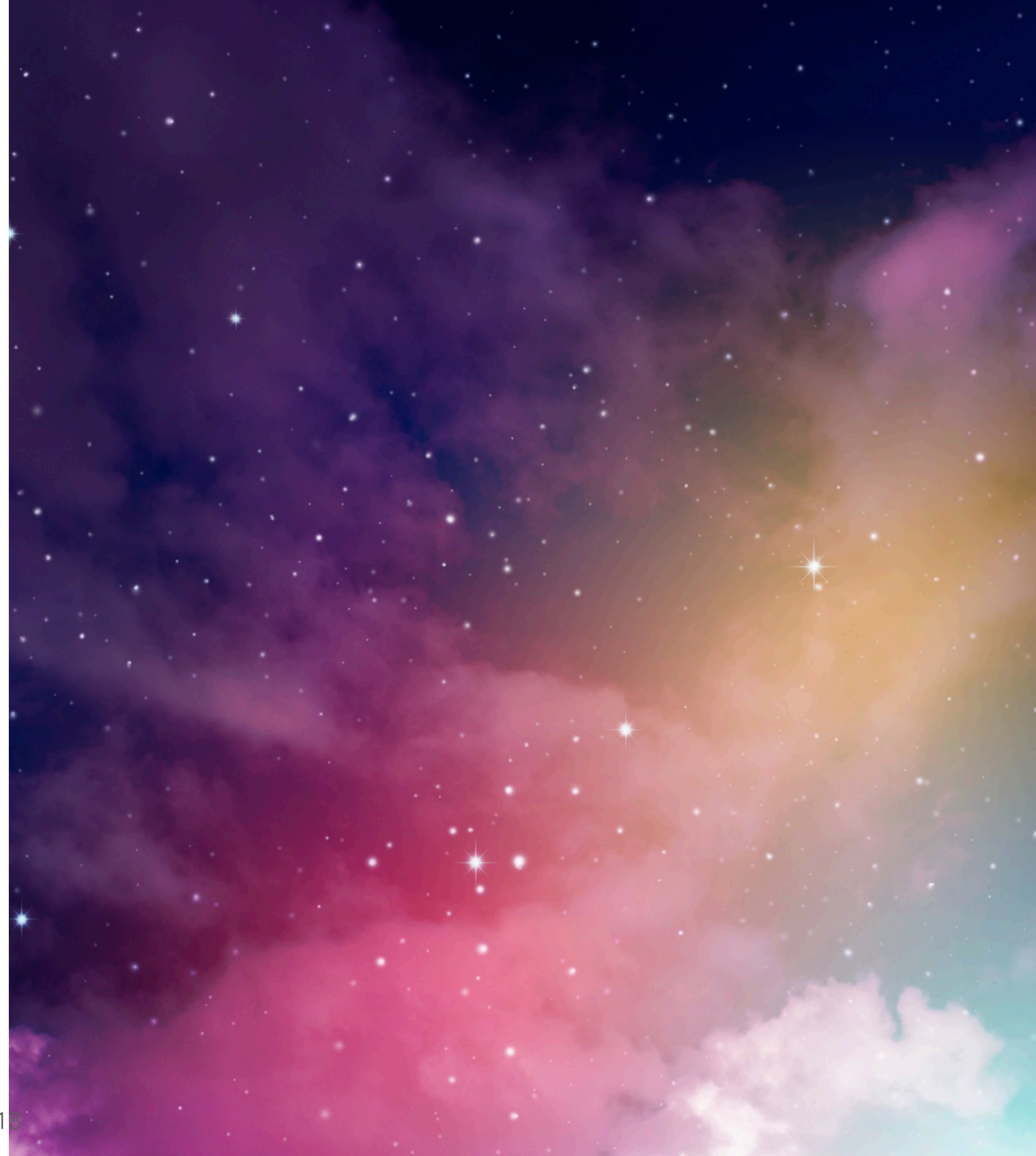
Number of conversations

Indicator if $s_{pc} = s_c$

Predicted speakers

True speakers

IV. Speaker Identification baseline

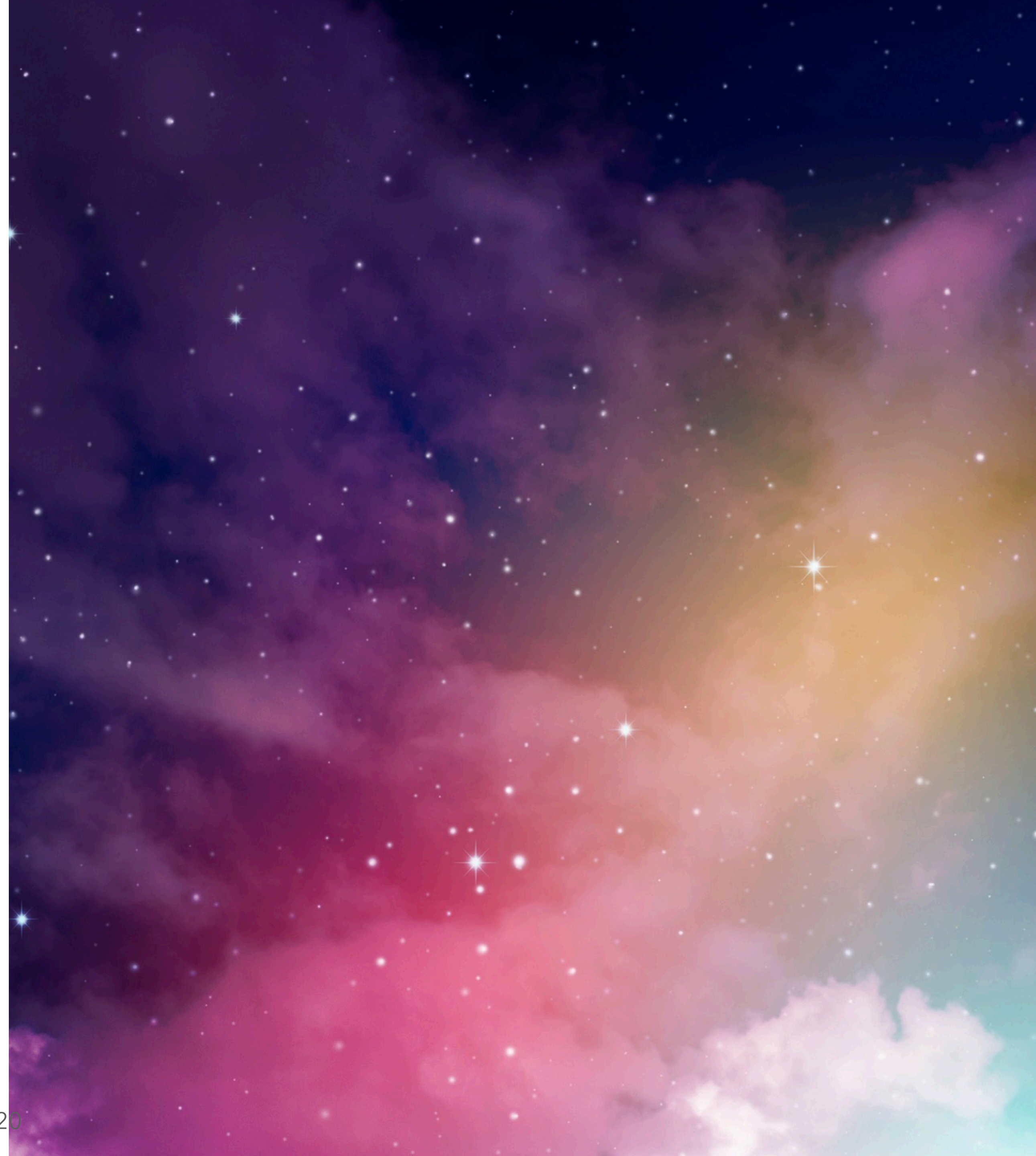


IV. Speaker identification baseline

Details of the system used:

- Used a pre-trained speaker identification system prepared for the NIST Speaker Recognition Evaluation (SRE19) dataset. (Idiap's submission to the NIST SRE 2019 Speaker Recognition Evaluation)
- Time Delay Neural Network (TDNN) X-vector systems with a PLDA back-end
- Downsampled speech data to 8kHz (with an application of band-pass filtering between 20 and 3700Hz)
- 23-dimensional MFCCs were extracted on 25ms speech windows, with a frame-shift of 10ms
- To remove non-speech frames, energy-based Voice Activity Detection (VAD) was applied

V. Re-ranking algorithm



V. Re-ranking algorithm

In CSI, many conversations involve more than 2 characters. And we only have 1 source of data. We need to bring solutions to the limits of previous approach.

$$s_{mc} = \frac{1}{N_{mc}} \left(\sum_{k=1}^{N_{mc}} s_k (1 + C_k) \right) S_m \left(1 + \lambda \frac{e_{k_1, k_2}}{E_c} \right)$$

Score of combination **m** in conversation **c**

Number of speakers in the conversation **c**

Acoustic score of speaker **k**

Degree centrality of speaker **k**

All 2*2 combinations of all speakers

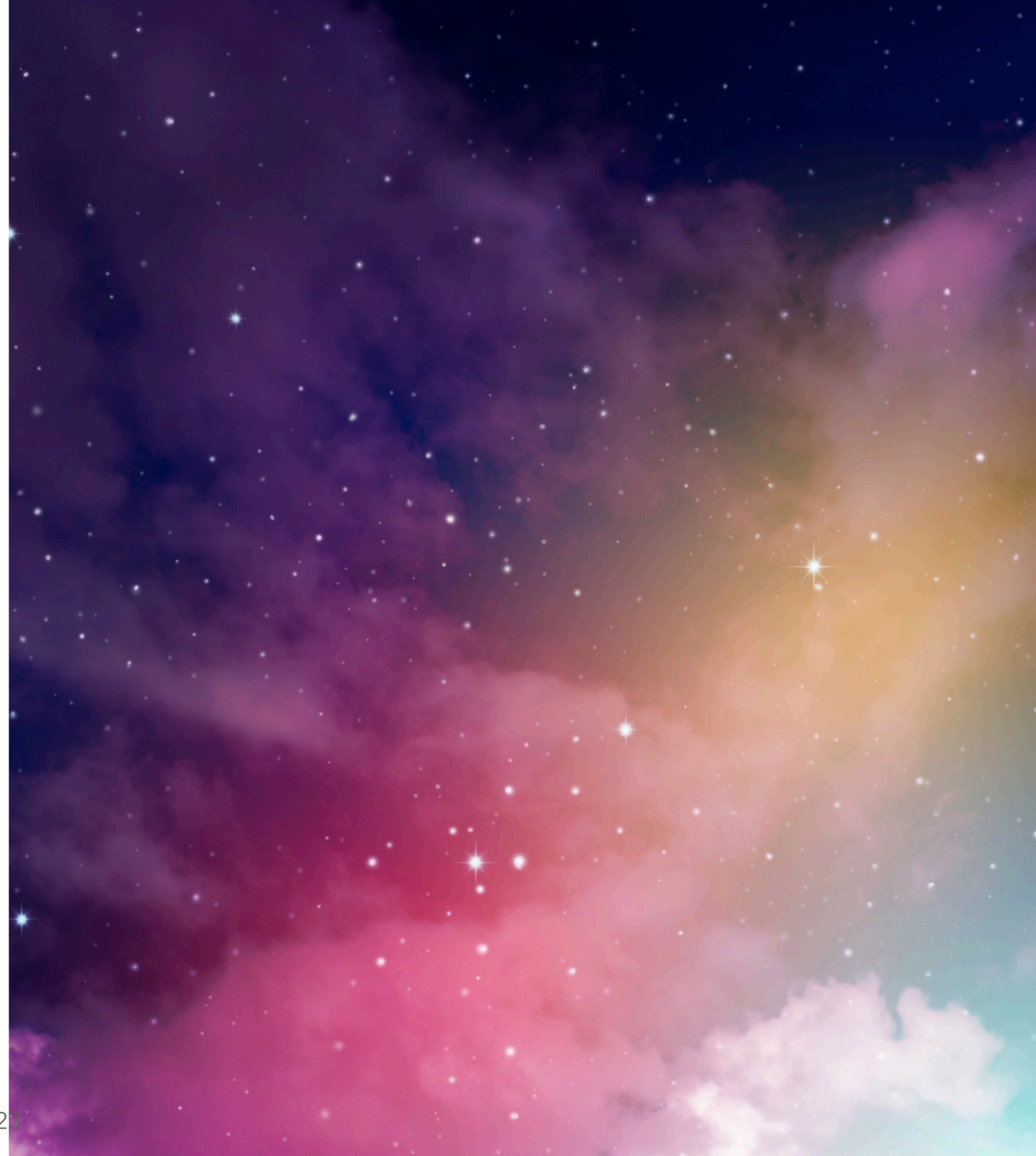
Percentage of conversations between 1 and 2 divided by total number of conversations

V. Re-ranking algorithm

We then select the combination that leads to the maximum score:

$$s_{mc}^* = \arg \max_{m \in M_c} s_{mc}$$

VI. Results



VI. Results

Results were extracted on 4 episodes on CSI.

We reached a relative improvement of 4.7% in terms of conversation accuracy and 1.5% in speaker accuracy.

For conversation and speaker accuracy, we obtained absolute improvements of 3.7% and 1.3%, respectively.

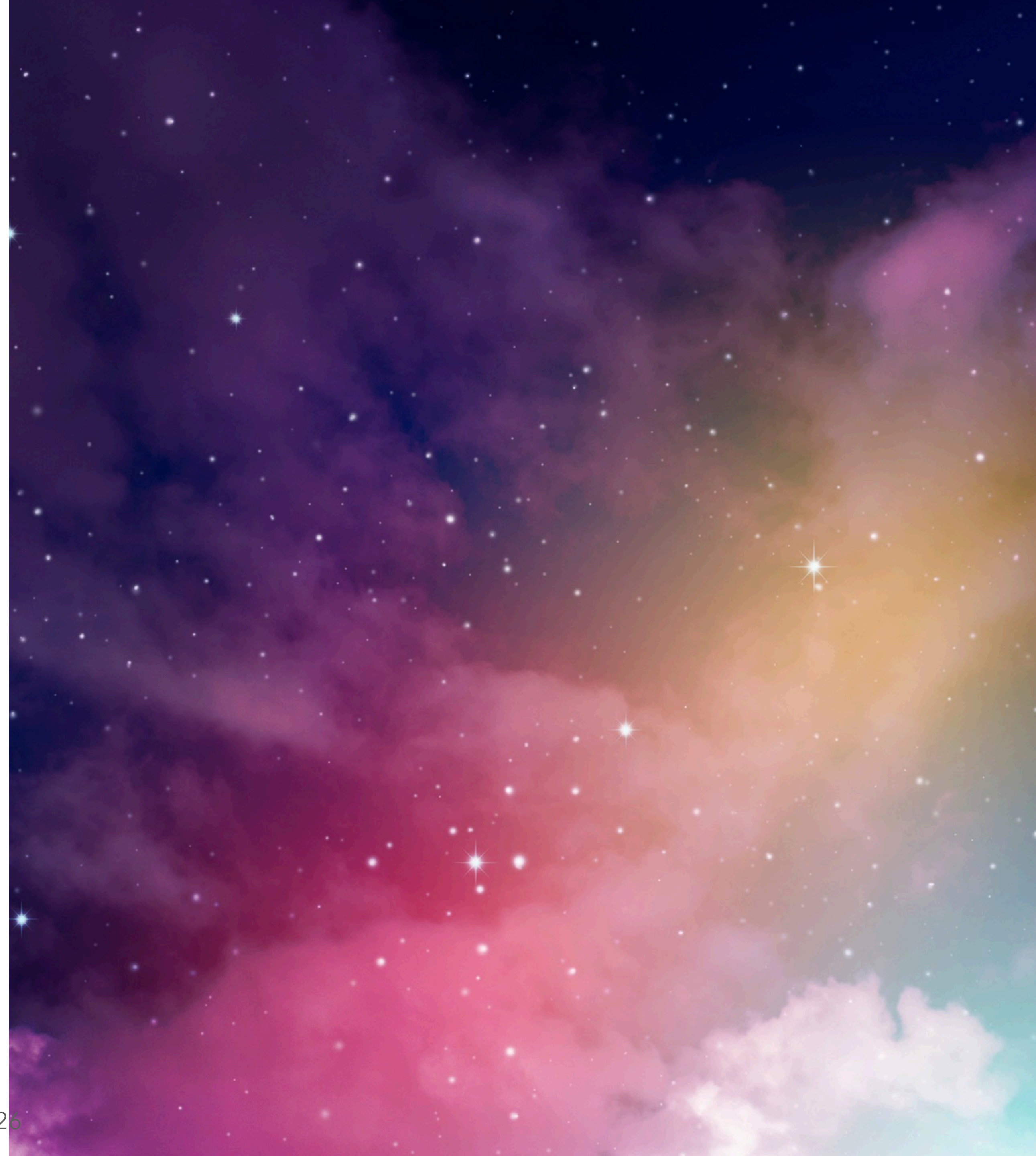
Approach	Speaker acc.	Conv. acc.
S01E07 baseline	91.6%	84.4%
S01E07 network	92.7%	88.8%
S01E08 baseline	91.9%	80.6%
S01E08 network	95.3%	88.8%
S02E01 baseline	88.0%	71.4%
S02E01 network	88.0%	73.5%
S02E04 baseline	88.1%	76.1%
S02E04 network	89.0%	76.1%
Average baseline	89.9%	78.1%
Average network	91.25%	81.8%

VI. Results

Performance is significantly improved when sub-groups are made at the beginning of the episode. However, when there is less structure in the investigation, the overall performance is less impacted.

VII.

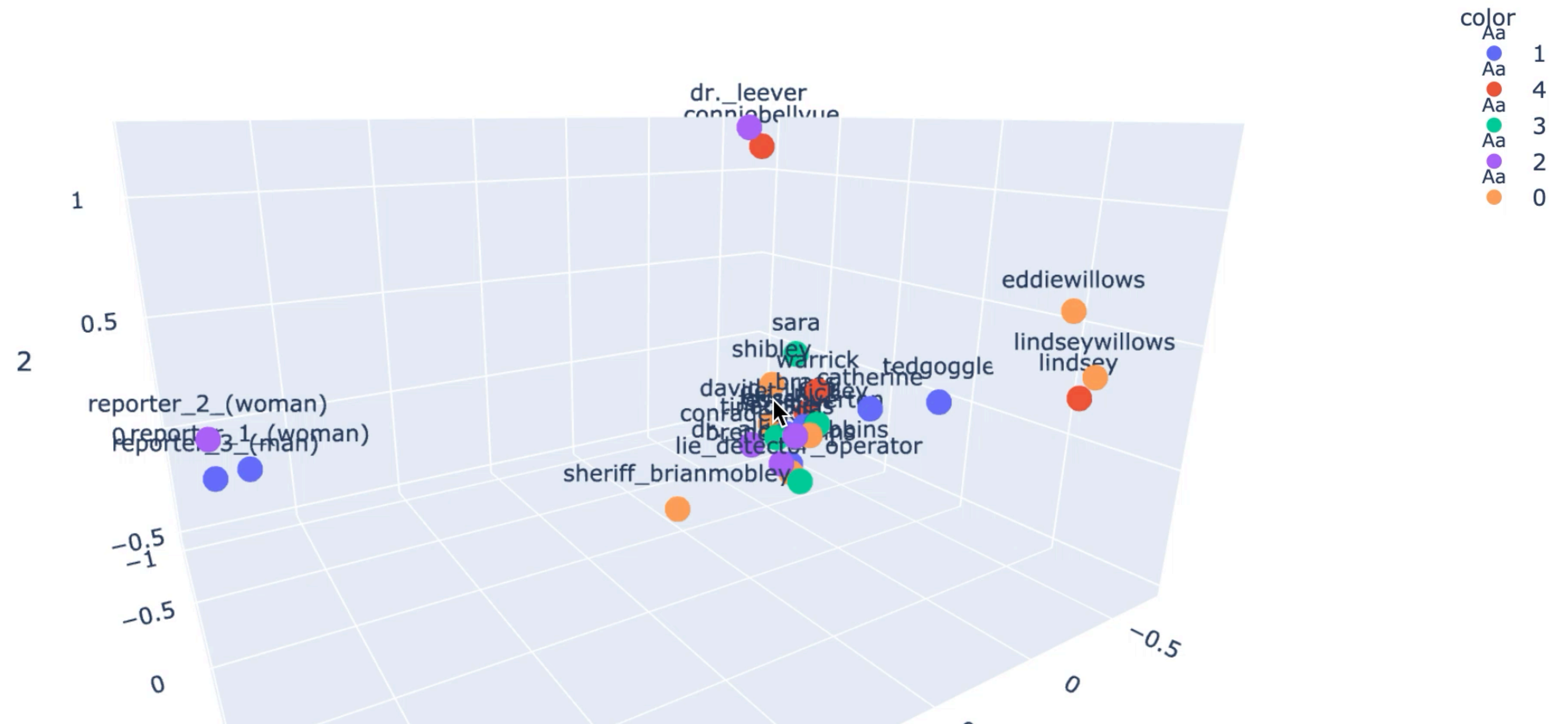
Future works



VII. Future works

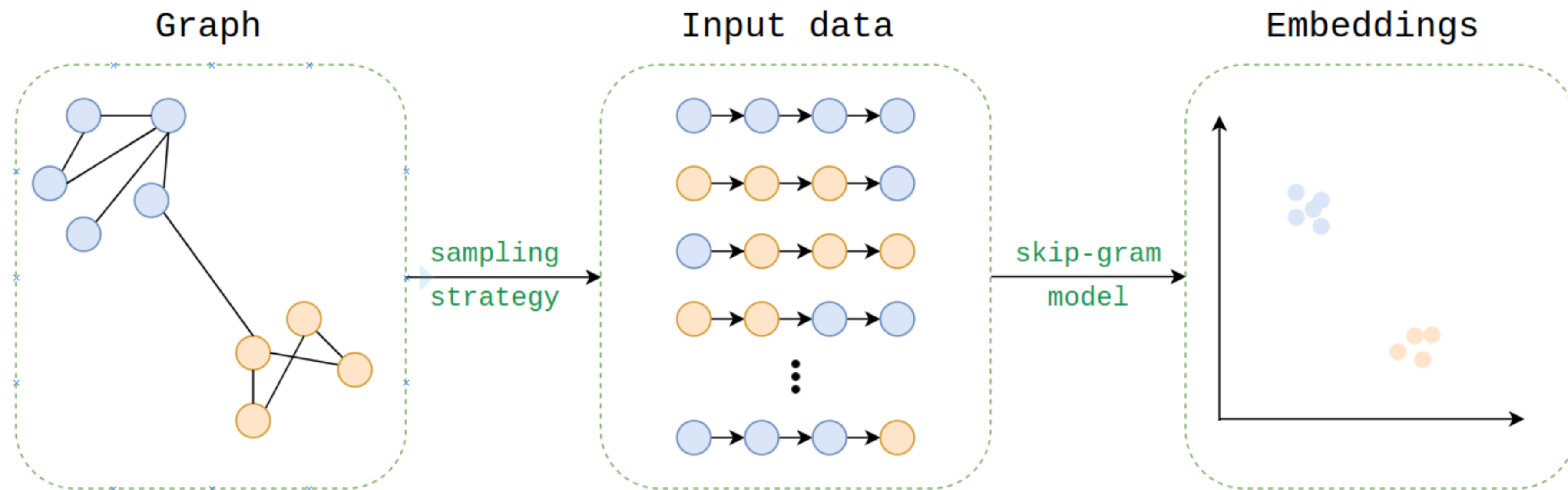
Re-ranking based on the number of edges between the potential characters is only one approach. We can explore several other approaches, including the similarity between node embeddings (Node2Vec).

VII. Future works



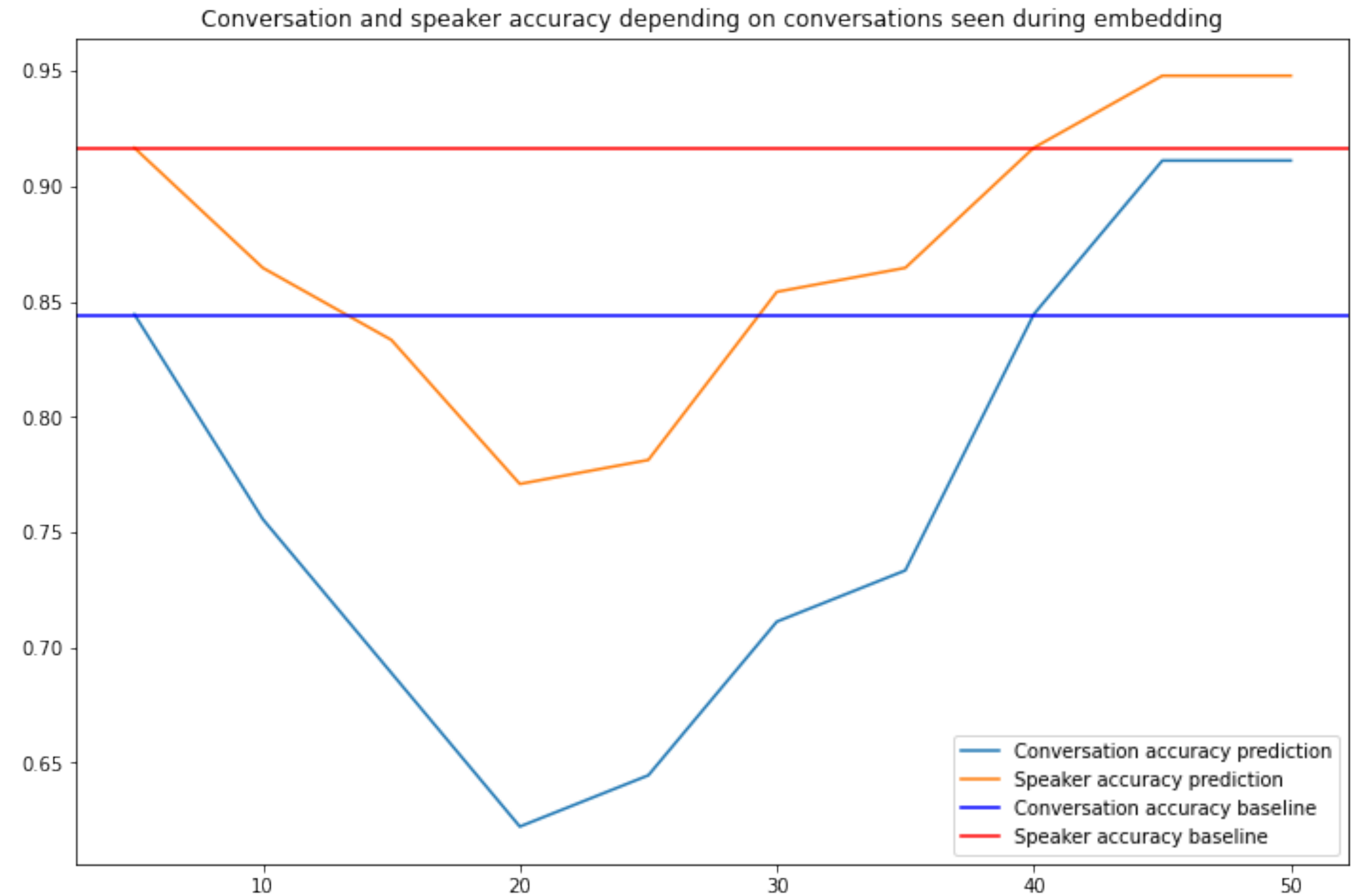
VII. Future works

We learn embeddings using Node2Vec, an implementation of Word2Vec on random walks in graphs.



VII. Future works

We must learn the embeddings on a relevant weighted graph. If the structure of the episode changes a lot after the embeddings we learned, the embeddings-based re-ranking won't improve the performance. But if the structure is relevant, we improve the performance by a significant factor.



Thank you for your attention
Questions?

Sources

- « Leveraging Side Information for Speaker Identification with the Enron Conversational Telephone Speech Collection », Ning Gao, Gregory Sell, Douglas W. Oard, Mark Dredze, http://www.cs.jhu.edu/~mdredze/publications/2017_asru_speakerid.pdf
- « Improving Speaker Identification using Network Knowledge in Criminal Conversational Data », Maël Fabien, Seyyed Saeed Sarfjoo, Petr Motlicek, Srikanth Madikeri, <https://arxiv.org/abs/2006.02093>