person instance **graphs** for person **recognition** in **multimedia** data

Hervé Bredin (CNRS / LIMSI)
Anindya Roy - Viet-Bac Le - Claude Barras

outline

- context
- **building** person instance graphs
- **mining** person instance graphs
- experimental **results**
- conclusion

the **REPERE** challenge

- multimodal person recognition in TV shows

- three ANR projects
  co-funded by DGA

QCOMPERE	**LIMSI** et al.
SODA	**LIUM** et al.
PERCOL	**LIF** et al.

- two evaluation campaigns in 2013 and 2014

  - data collection & annotation by ELDA
  - evaluation by LNE

multimodal person recognition

- **multiple** sources of information

  - **audio stream**
    speaker diarization & identification
    speech transcription (ASR)

  - **visual stream**
    face clustering & recognition
    optical character recognition (OCR)

  - **text stream** (from ASR and OCR)
    named entity detection
    name normalization

the QCompere project

BY quaero

the **REPERE** challenge

who speaks when?

who appears when?

the **REPERE** challenge

who speaks when?

who appears when?

the **holy grail**
free annotated data

- **dense audio annotation**

  speaker identification
  speech transcription

- **sparse video annotation**

  head recognition
  optical character recognition

- **text annotation** named entity detection (person names only)

the **holy grail**
free annotated data

- **dense audio annotation**

  speaker identification
  speech transcription

- **sparse video annotation**

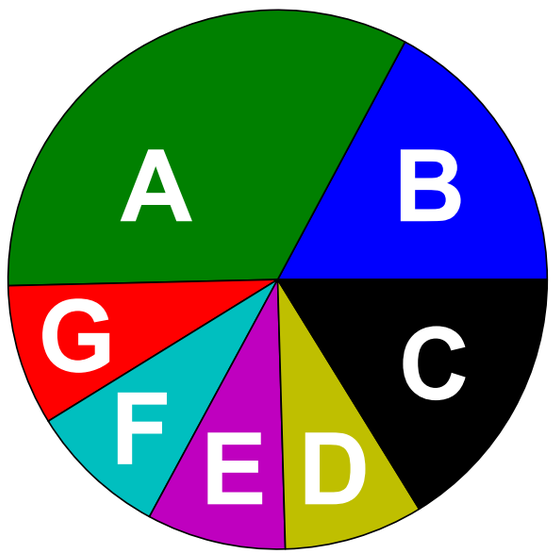
  head recognition
  optical character recognition

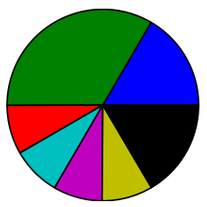
- **text annotation** named entity detection (person names only)

2013

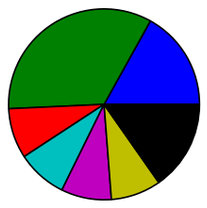
TRAINING SET
24 hours

A: BFM Story
B: LCP Info
C: Top Questions
D: Ça Vous Regarde
E: Planète Showbiz
F: Entre Les Lignes
G: Pile Et Face





DEVELOPMENT
3 hours



TEST
3 hours

the **holy grail**
free annotated data

- **dense audio annotation**

  speaker identification
  speech transcription

- **sparse video annotation**

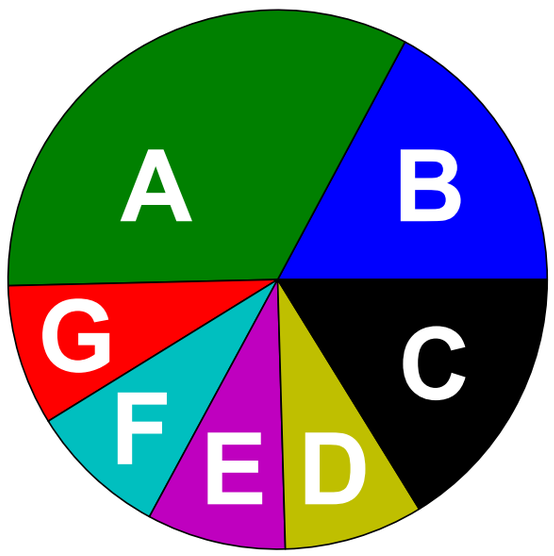
  head recognition
  optical character recognition

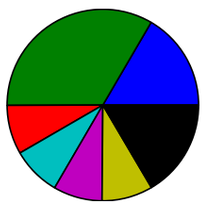
- **text annotation** named entity detection (person names only)

2013

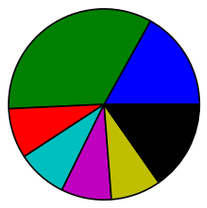
TRAINING SET
24 hours

A: BFM Story
B: LCP Info
C: Top Questions
D: Ça Vous Regarde
E: Planète Showbiz
F: Entre Les Lignes
G: Pile Et Face





DEVELOPMENT
3 hours



TEST
3 hours

2014

X 2

various **subtasks**

- **supervised** recognition
  can rely on prior biometric models (face or voice) from manually annotated data
  can rely on written and/or pronounced person names

- **unsupervised** recognition
  must not rely on prior biometric models
  must rely on written and/or pronounced person names

- **mono-modal** recognition
  use audio signal only for speaker recognition
  use facial features for face recognition

- **cross-modal** recognition
  use written/spoken name for speaker recognition

- **multi-modal** recognition
  use anything

why is this a **difficult task**?

- **missing** training data
  35% of test speakers have no training data
  even worse for face recognition

- **asynchronous** data
  a speaker may be heard but not seen
  two (or more) people may appear at the same time
  people rarely pronounces their own name

  simple score fusion is not an option

- **heterogeneous** data
  voice, face, text from OCR and from ASR

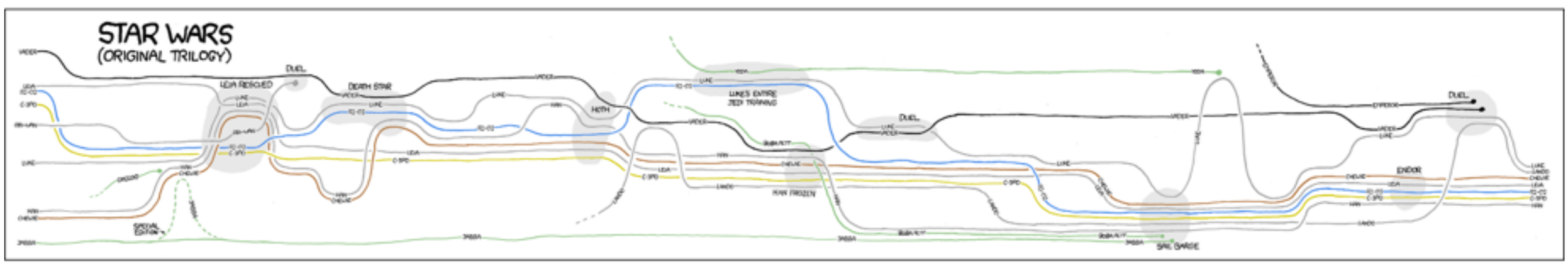
  needs expertise in speech processing, computer vision and machine learning

why is this a **useful task**?

- **information retrieval**
  find all sequences where Sarkozy talks about immigration

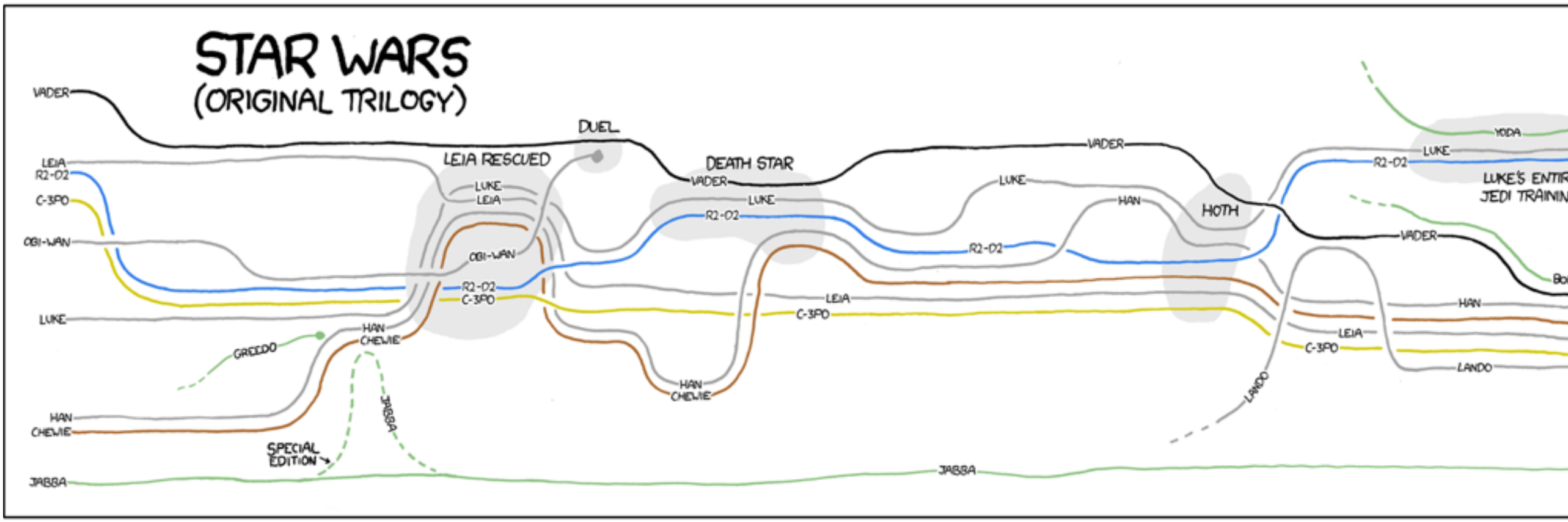
- **radio/TV broadcast monitoring**
  measure total speech time for every politician (CSA)

- **second-screen** applications
  narrative graphs of TV/movie series



xkcd.com/657/

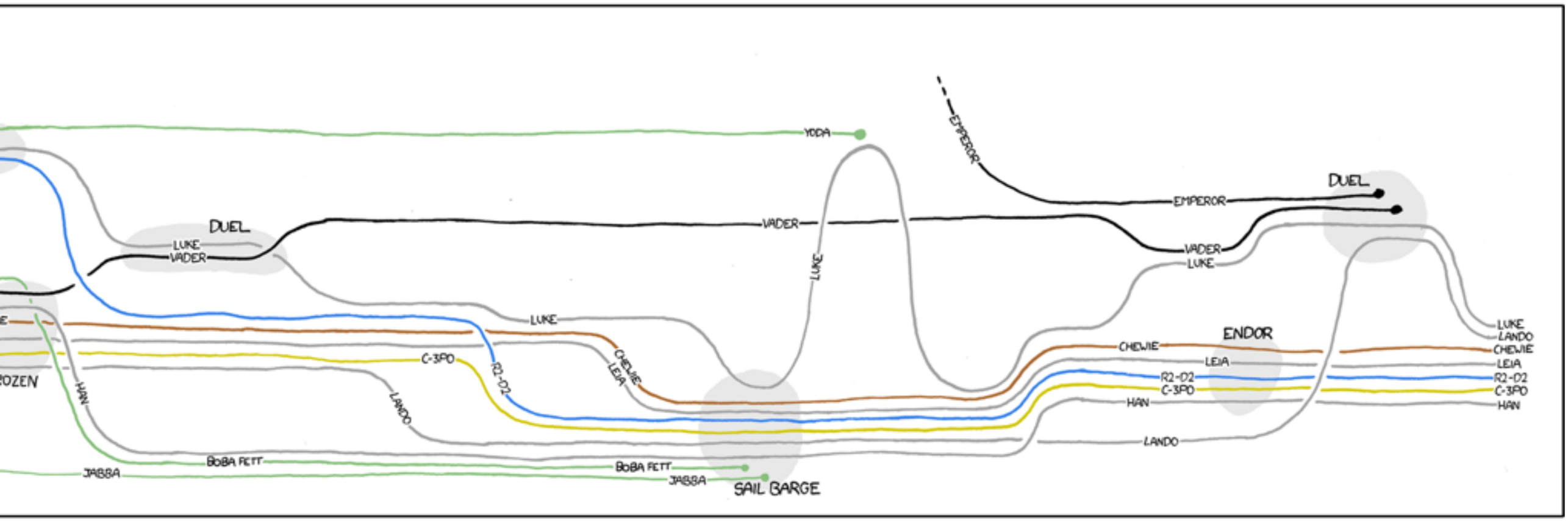
narrative graphs of TV/movie series



xkcd.com/657/

Makarand Tapaswi, Martin Baeuml and Rainer Stiefelhagen
StoryGraphs: Visualizing Character Interactions as a Timeline
IEEE Computer Vision for Pattern Recognition, Columbus, June 2014

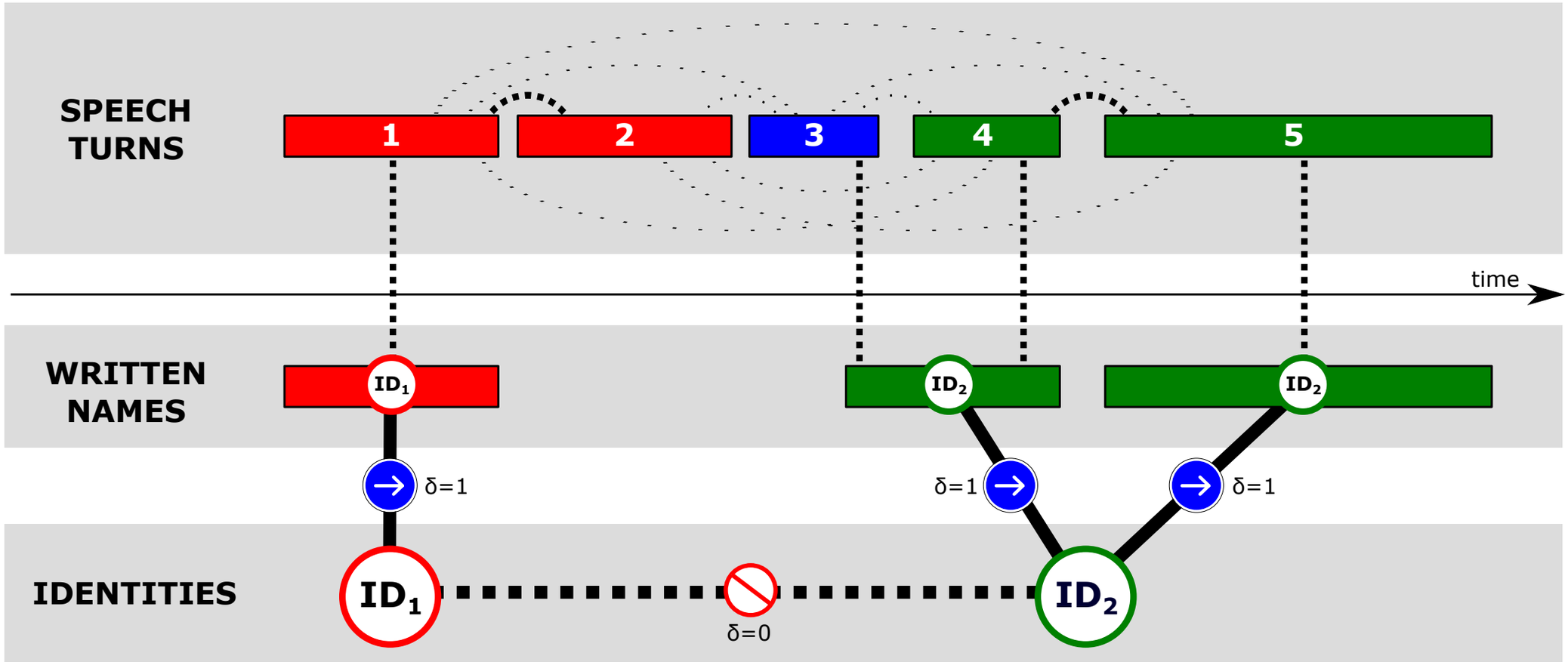
narrative graphs of TV/movie series



xkcd.com/657/

Makarand Tapaswi, Martin Baeuml and Rainer Stiefelhagen
StoryGraphs: Visualizing Character Interactions as a Timeline
IEEE Computer Vision for Pattern Recognition, Columbus, June 2014

building
person instance graphs



person instance **graph**

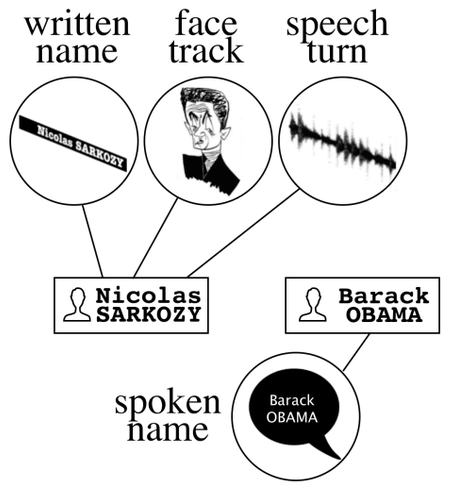
- weighted undirected graph
  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, p)$

- instance vertices

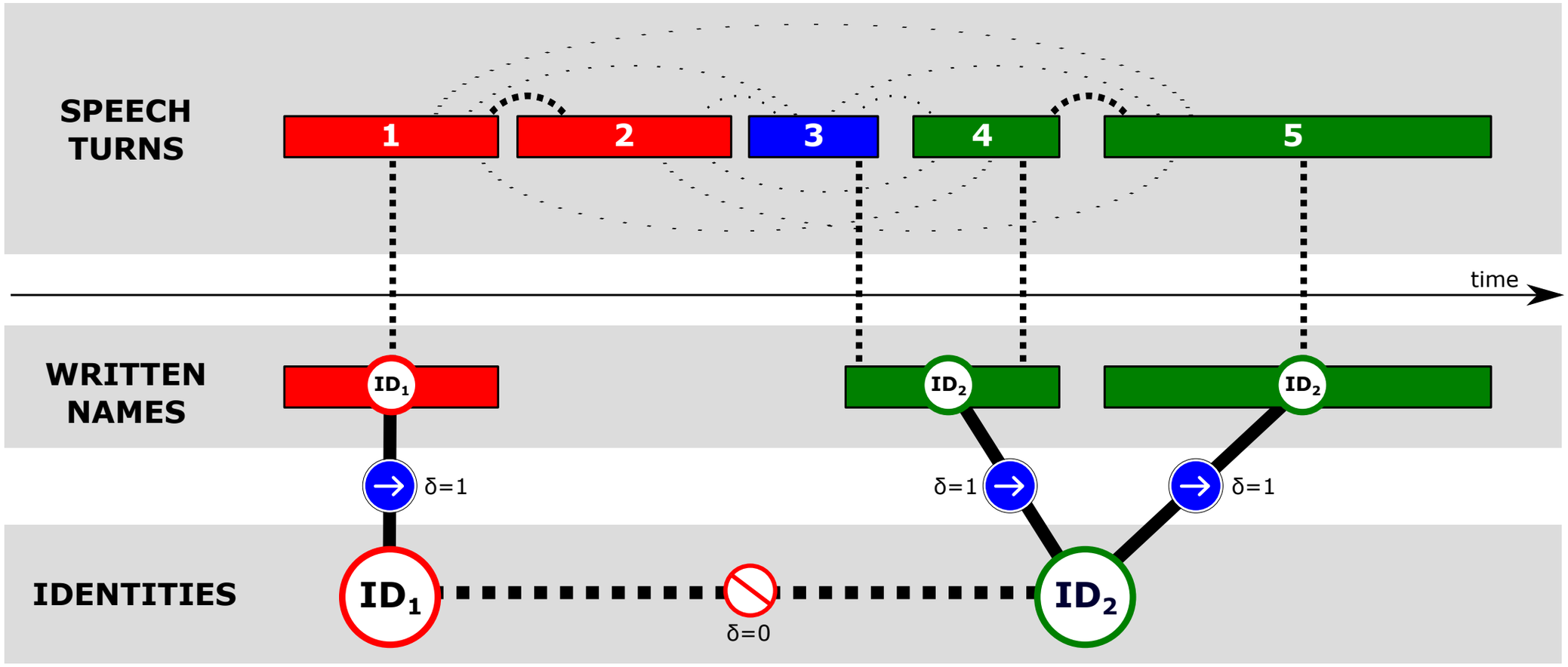
  speech turns $t \in \mathcal{T}$
  written names $w \in \mathcal{W}$
  spoken names $s \in \mathcal{S}$
  ~~face tracks~~

- identity vertices

- edges $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$
  weighted by $p \in [0, 1]^{\mathcal{E}}$

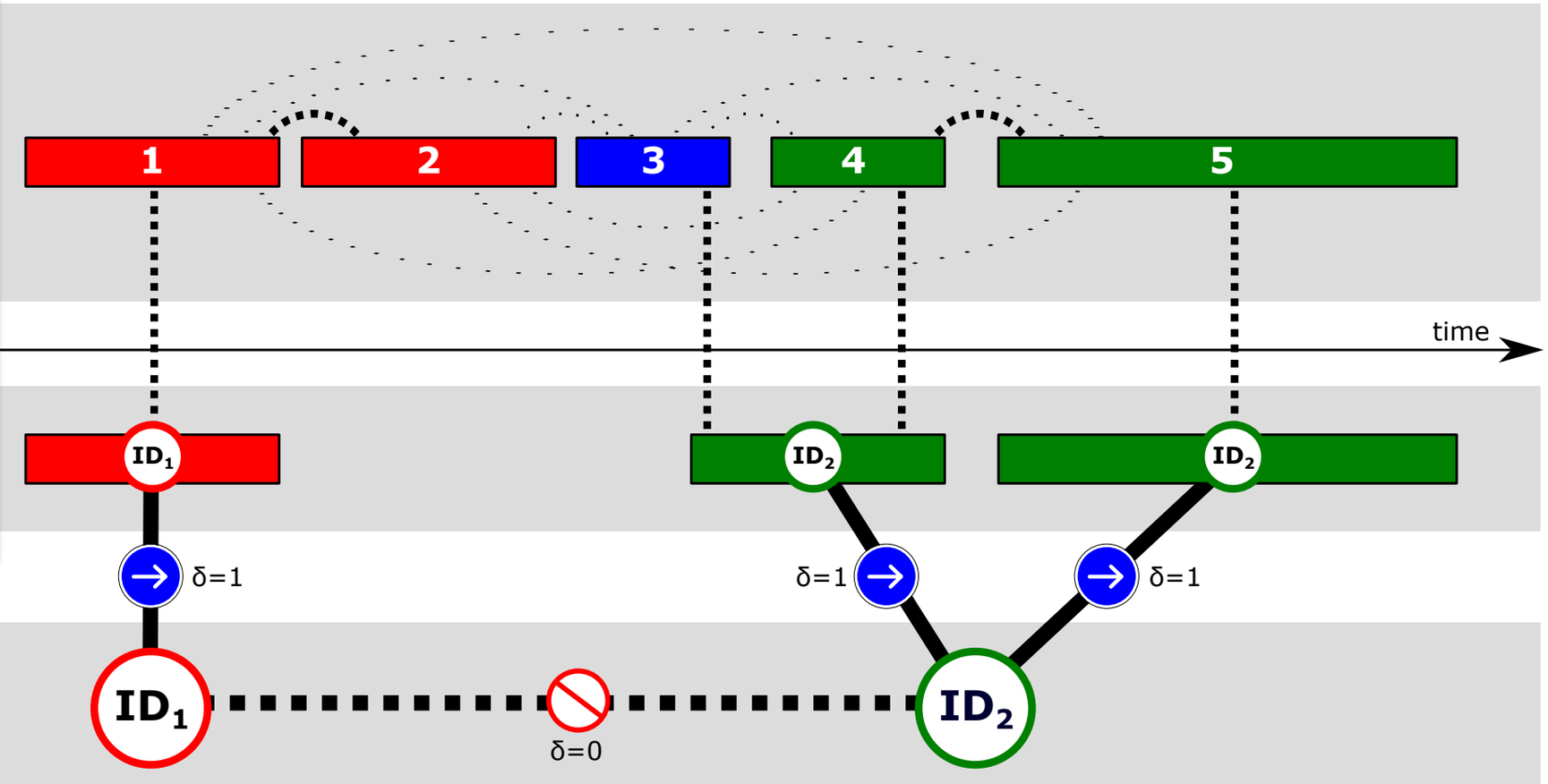


vertices


SPEECH TURNS
1
2
3
4
5
time
WRITTEN NAMES
ID1
ID2
ID2
δ=1
δ=1
δ=1
IDENTITIES
ID1
δ=0
ID2

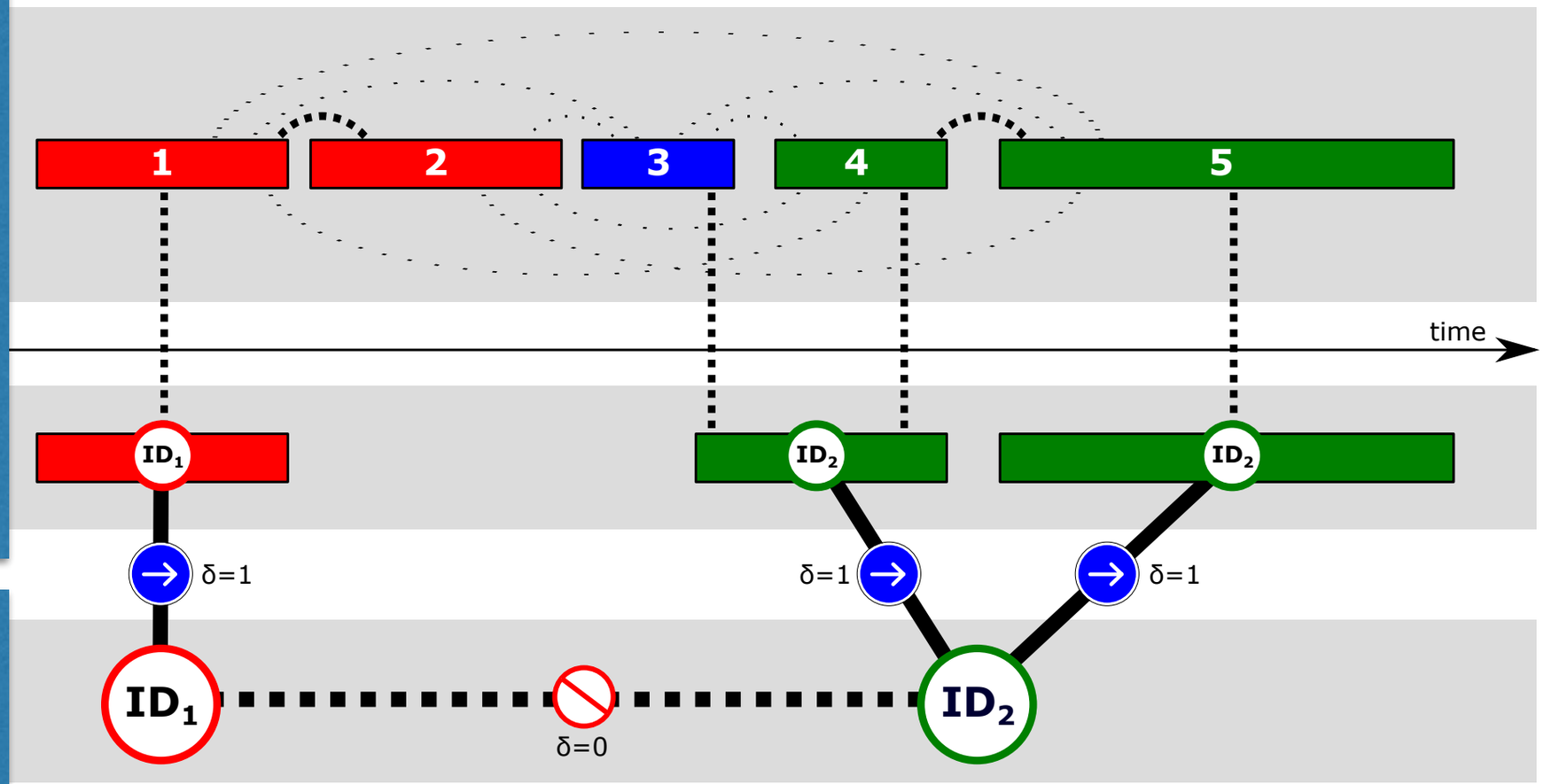
vertices

instance
vertices



vertices

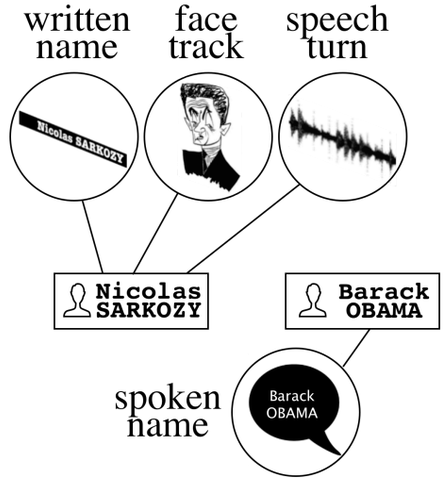
instance vertices

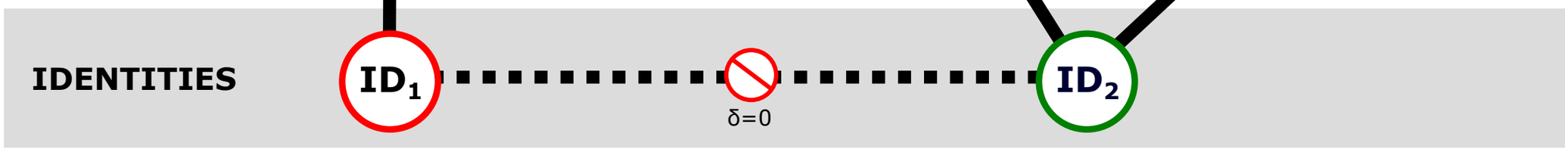


identity vertices

instance vertices

- **four** potential **sources of instance** vertices
  - **speech turns** $t \in \mathcal{T}$
    speech activity detection
    divergence-based segmentation
  - **face tracks** (not used here)
    face detection and tracking
  - **written names** $w \in \mathcal{W}$
    video optical character recognition
    named entity detection
  - **spoken names** $s \in \mathcal{S}$
    manual speech transcription
    manual named entity detection





identity vertices

- **four** potential **sources of identity** vertices

  one unique identifier per person
  Nicolas Sarkozy ⇒ Nicolas_SARKOZY

  - **one identity vertex** per prior **speaker model** $\mathcal{I}^*$
    350 speaker models ⇒ 350 identity vertices

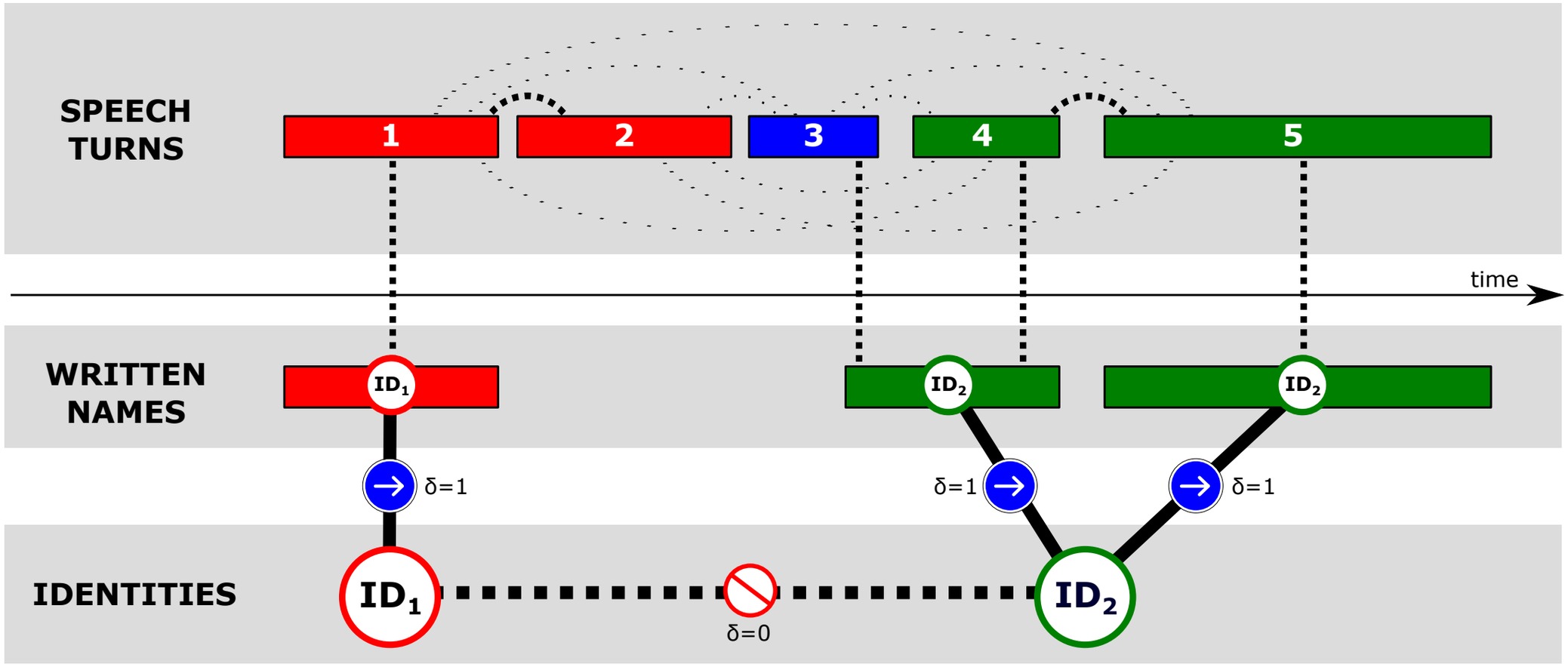
  - **one identity vertex** per prior **face model**
    not used here

  - **one identity vertex** for each **written name transcription** $\mathcal{I}_\mathcal{W}$

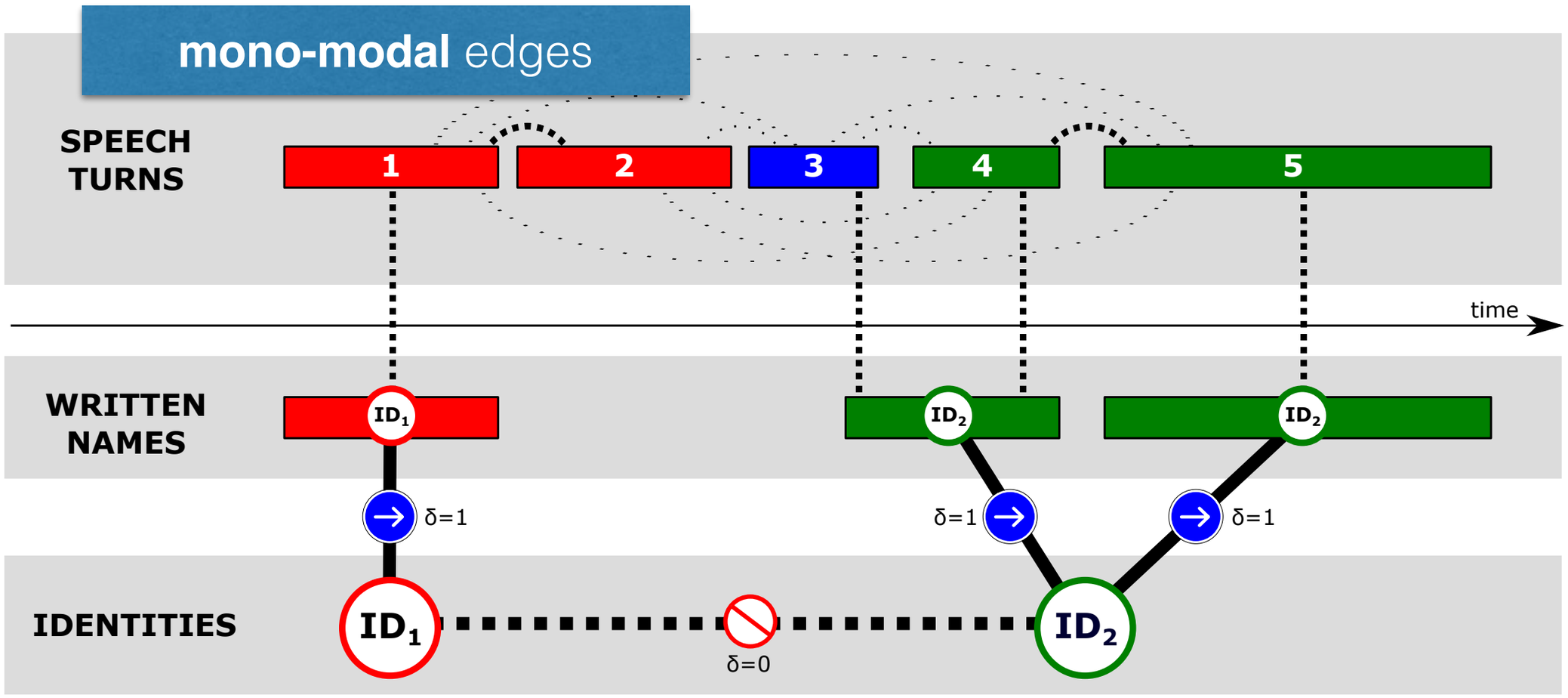
  - **one identity vertex** for each **spoken name transcription** $\mathcal{I}_\mathcal{S}$

- $\mathcal{I} = \mathcal{I}_\mathcal{W} \cup \mathcal{I}_\mathcal{S} \cup \mathcal{I}^*$
  an identity vertex can belong to more than one subset…

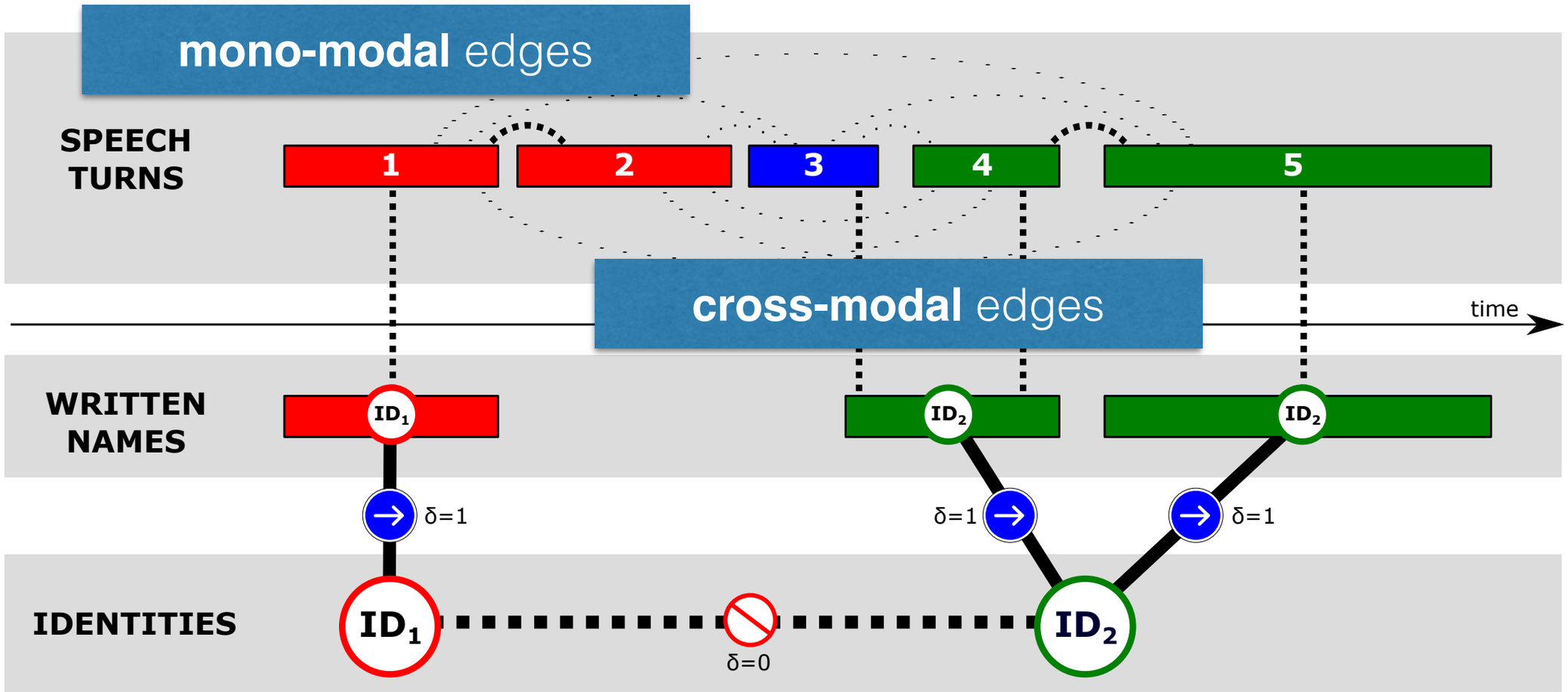
edges


SPEECH TURNS
1
2
3
4
5
time
WRITTEN NAMES
ID1
ID2
ID2
δ=1
δ=1
δ=1
IDENTITIES
ID1
δ=0
ID2

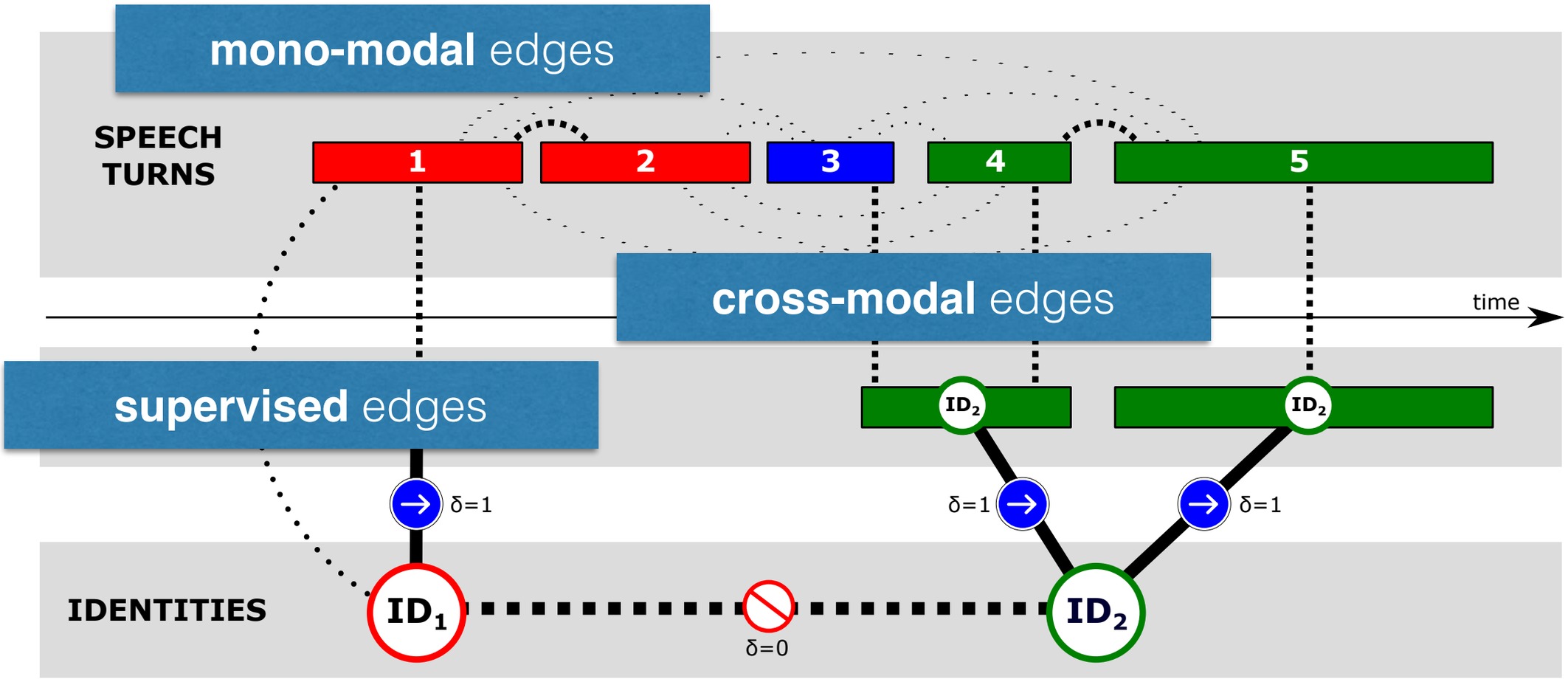
edges


mono-modal edges
SPEECH TURNS
1
2
3
4
5
time
WRITTEN NAMES
ID1
ID2
ID2
δ=1
δ=1
δ=1
IDENTITIES
ID1
δ=0
ID2

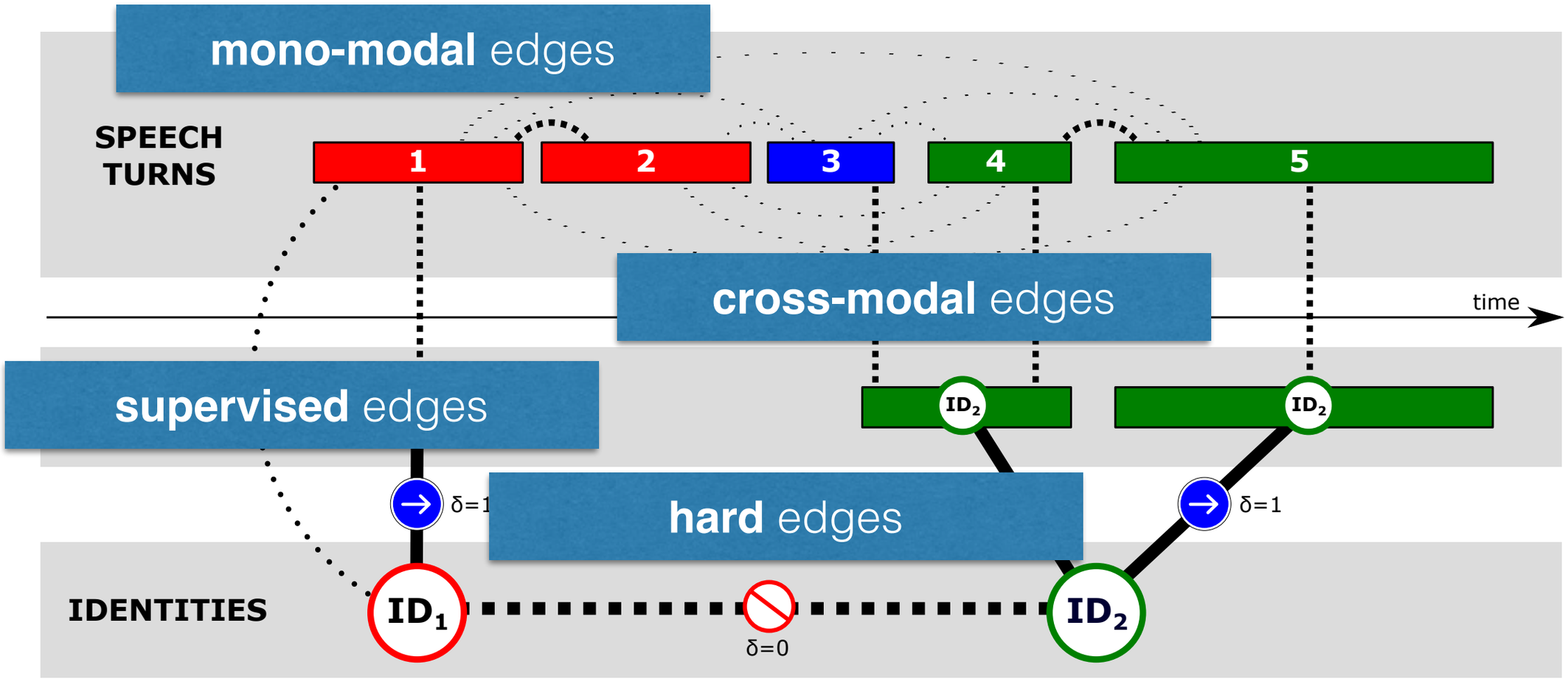
edges


mono-modal edges
SPEECH TURNS
1
2
3
4
5
cross-modal edges
time
WRITTEN NAMES
ID1
ID2
ID2
δ=1
δ=1
δ=1
IDENTITIES
ID1
δ=0
ID2

edges


mono-modal edges
SPEECH TURNS
1
2
3
4
5
cross-modal edges
time
supervised edges
ID2
ID2
δ=1
δ=1
δ=1
IDENTITIES
ID1
δ=0
ID2

edges

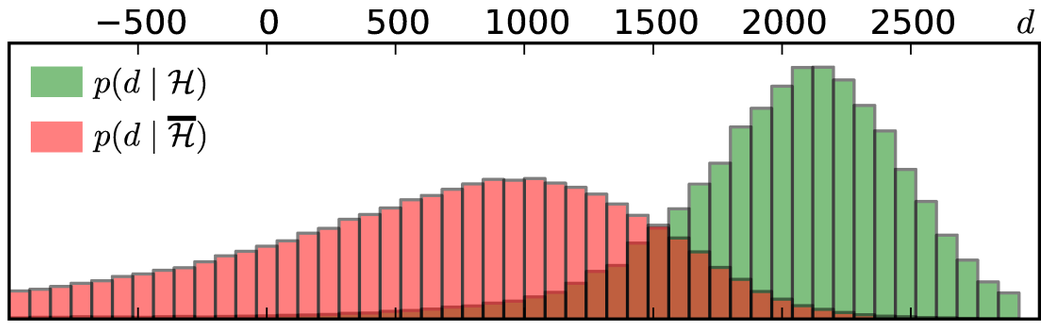

mono-modal edges
SPEECH TURNS
1
2
3
4
5
cross-modal edges
time
supervised edges
ID2
ID2
δ=1
hard edges
δ=1
IDENTITIES
ID1
δ=0
ID2

mono-modal edges

(similarity between **speech turns**)

- Standard MFCC features
- Bayesian information criterion (BIC)

$$
\begin{aligned}
d_{tt'} &= (n_t + n_{t'}) \log |\Sigma_{t+t'}| \\
&\quad - n_t \log |\Sigma_t| - n_{t'} \log |\Sigma_{t'}| \\
&\quad - \frac{1}{2} \cdot \lambda \cdot \left(D + \frac{1}{2} D \left(D + 1 \right) \right) \log \left(n_t + n_{t'} \right)
\end{aligned}
$$

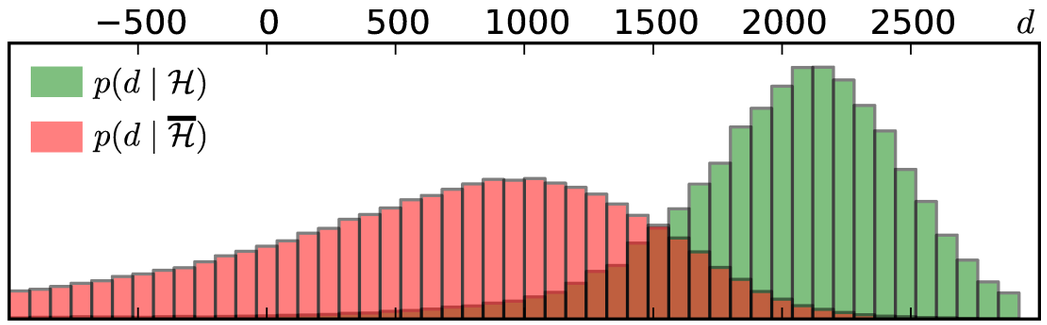


mono-modal edges

(similarity between **speech turns**)

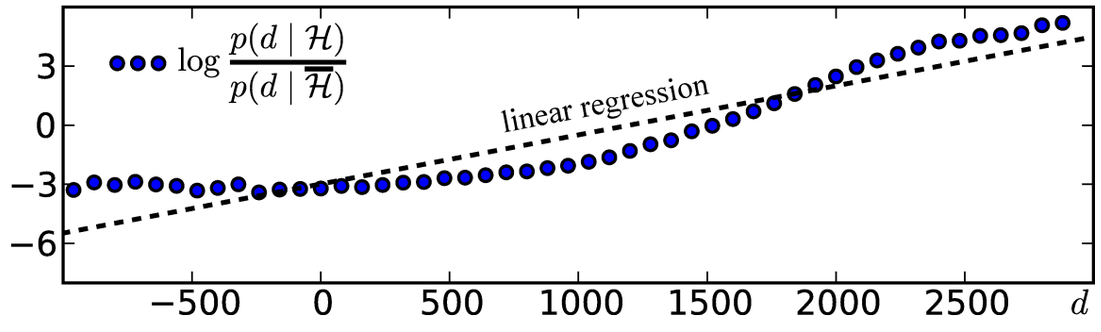
- Standard MFCC features
- Bayesian information criterion (BIC)

$$
\begin{aligned}
d_{tt'} &= (n_t + n_{t'}) \log |\Sigma_{t+t'}| \\
&\quad - n_t \log |\Sigma_t| - n_{t'} \log |\Sigma_{t'}| \\
&\quad - \frac{1}{2} \cdot \lambda \cdot \left(D + \frac{1}{2} D \left(D + 1 \right) \right) \log \left(n_t + n_{t'} \right)
\end{aligned}
$$

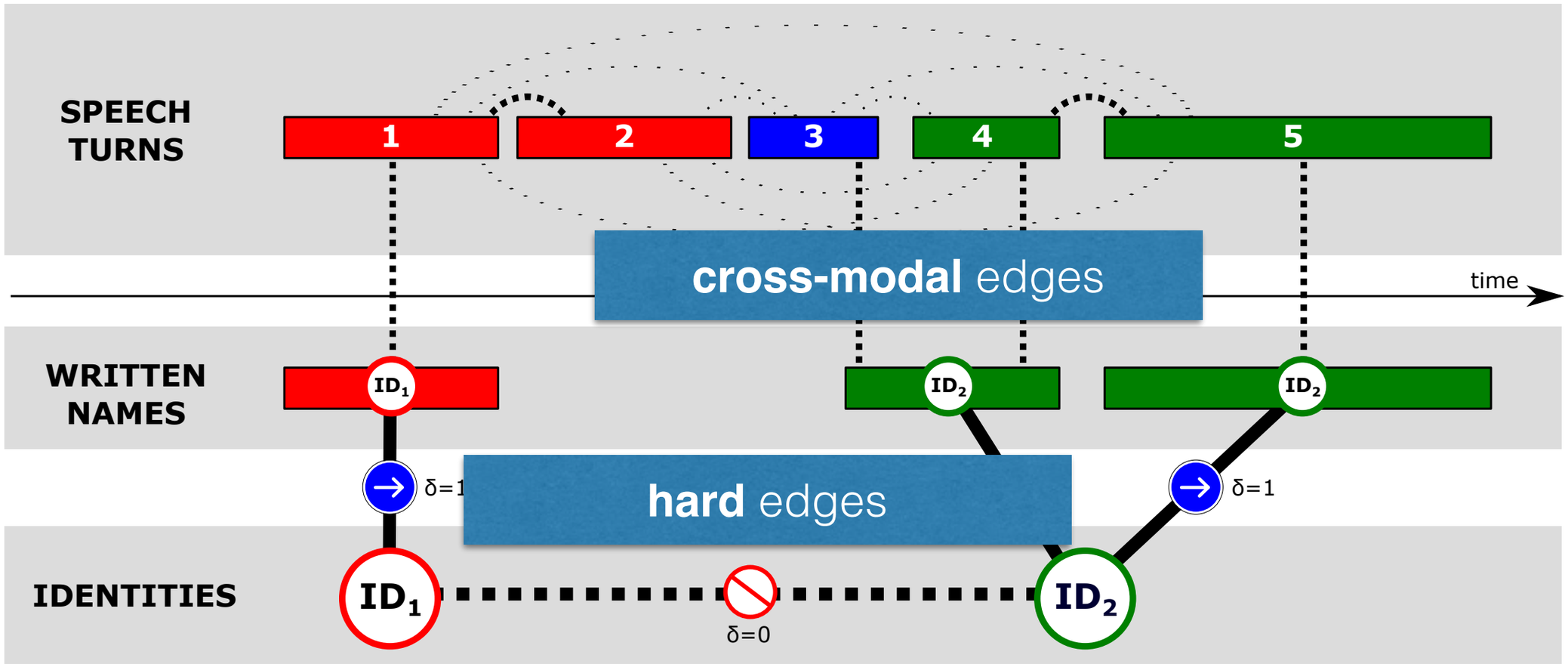


- Bayes' theorem
- Linear regression for log-likelihood ratio estimation

$$
\begin{aligned}
p_{tt'} &= p(\texttt{ID}(t) = \texttt{ID}(t') \mid d_{tt'}) \\
&= \frac{1}{1 + \dfrac{\pi_{\neq}}{\pi_{=}} \cdot \dfrac{p(d_{tt'} \mid \texttt{ID}(t) \neq \texttt{ID}(t'))}{p(d_{tt'} \mid \texttt{ID}(t) = \texttt{ID}(t'))}}
\end{aligned}
$$



edges


SPEECH TURNS
1
2
3
4
5
time
cross-modal edges
WRITTEN NAMES
ID1
ID2
ID2
δ=1
hard edges
δ=1
IDENTITIES
ID1
δ=0
ID2

cross-modal edges

$t \leftrightarrow w$

probability that a name written at time t
is the name of current speaker

$p_{tw} = 0.956$

$p_{tw} = 0.996$

cross-modal edges

$t \leftrightarrow w$

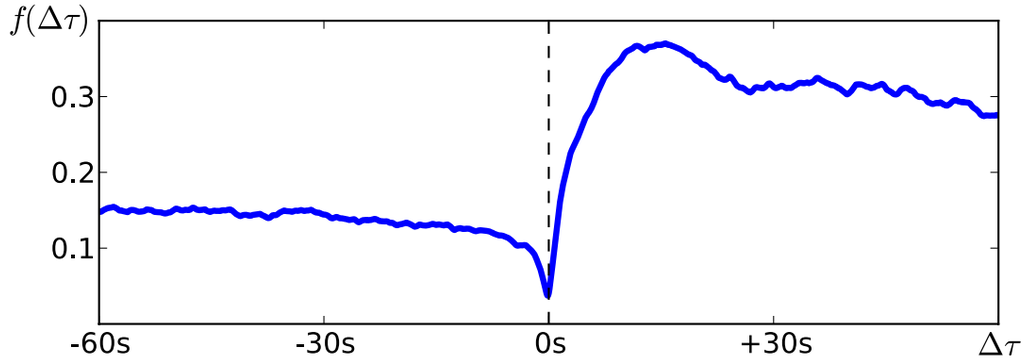
probability that a name written at time t
is the name of current speaker

$p_{tw} = 0.956$

$p_{tw} = 0.996$

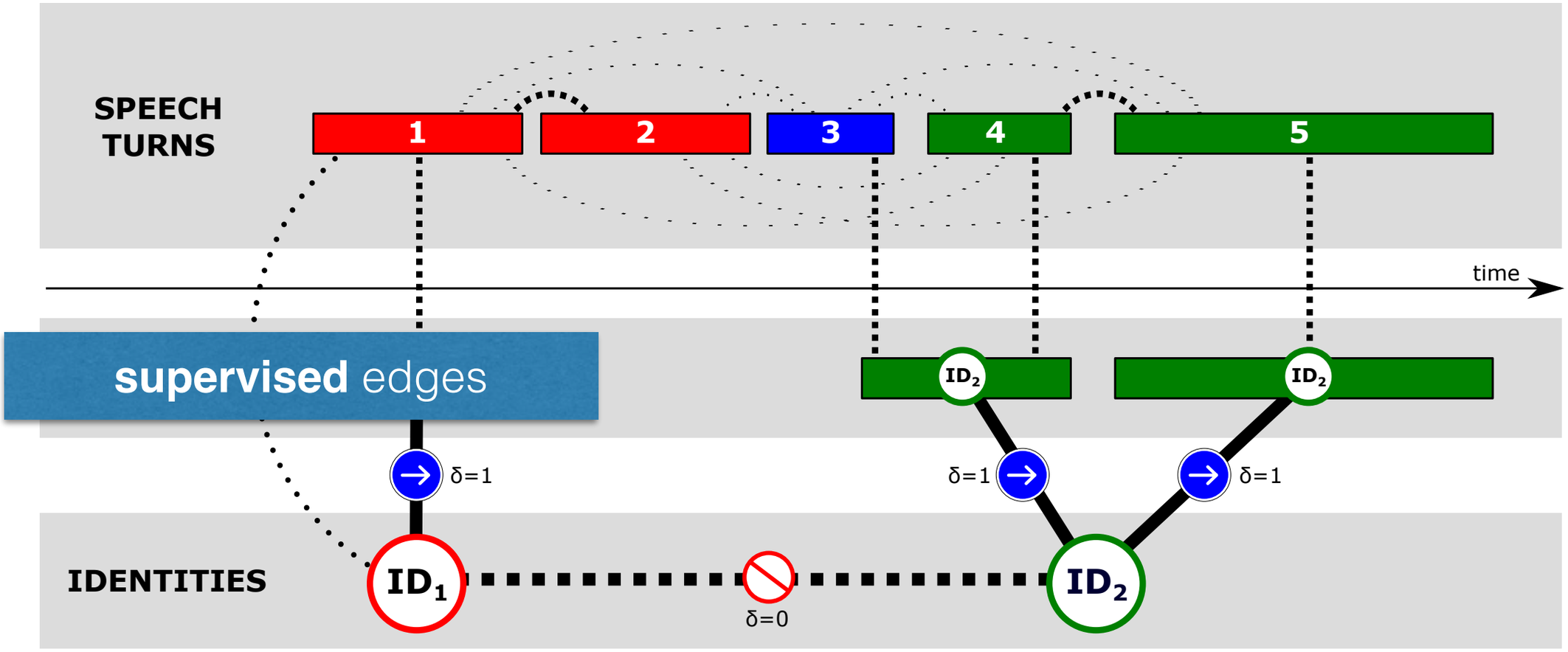
$t \leftrightarrow s$

probability that a name pronounced at time t
is the name of speaker at time t + dt



$$p_{ts} = p(\texttt{ID}(t) = \texttt{ID}(s) \mid t, s)$$
$$= \frac{1}{|T(t)|} \int_{\tau \in T(t)} f(\tau - \tau_s)\, \mathrm{d}\tau$$

edges


SPEECH TURNS
1
2
3
4
5
time
supervised edges
ID2
ID2
δ=1
δ=1
δ=1
IDENTITIES
ID1
δ=0
ID2

supervised edges

- GMM / UBM speaker identification

$$d_{ti} = \frac{1}{n_t}\left[\log \prod_{x \in X_t} f\left(x|\lambda_i\right) - \log \prod_{x \in X_t} f\left(x|\lambda_\Omega\right)\right]$$

$$p_{ti} = p(\texttt{ID}(t) = i \mid d_{ti})$$

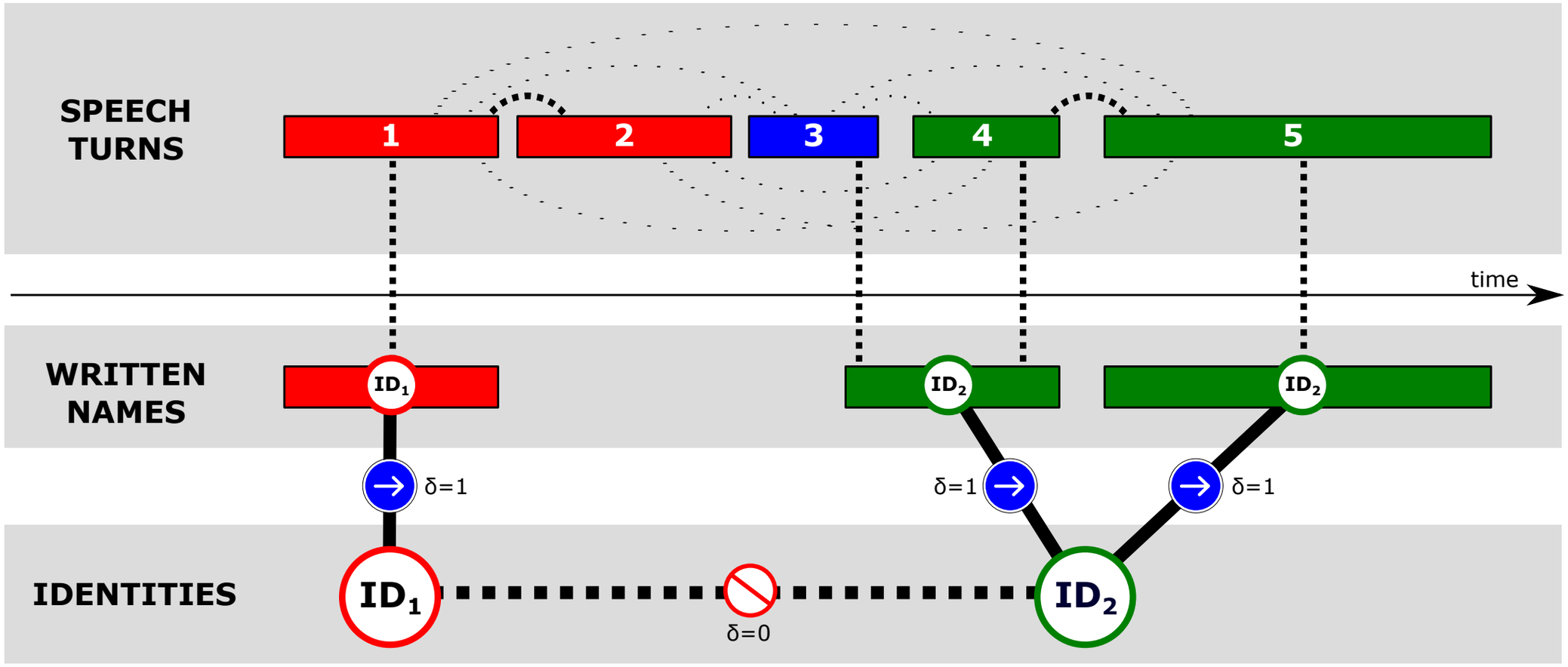
$$= \frac{\pi_i \cdot \dfrac{p(d_{ti} \mid \texttt{ID}(t) = i)}{p(d_{ti} \mid \texttt{ID}(t) \neq i)}}{\pi_{\textcircled{?}} + \displaystyle\sum_{i' \in \mathcal{I}^*} \pi_{i'} \cdot \frac{p(d_{ti'} \mid \texttt{ID}(t) = i')}{p(d_{ti'} \mid \texttt{ID}(t) \neq i')}}$$

- every speech turn is connected $t \leftrightarrow i$ to all (≈350) identity vertices

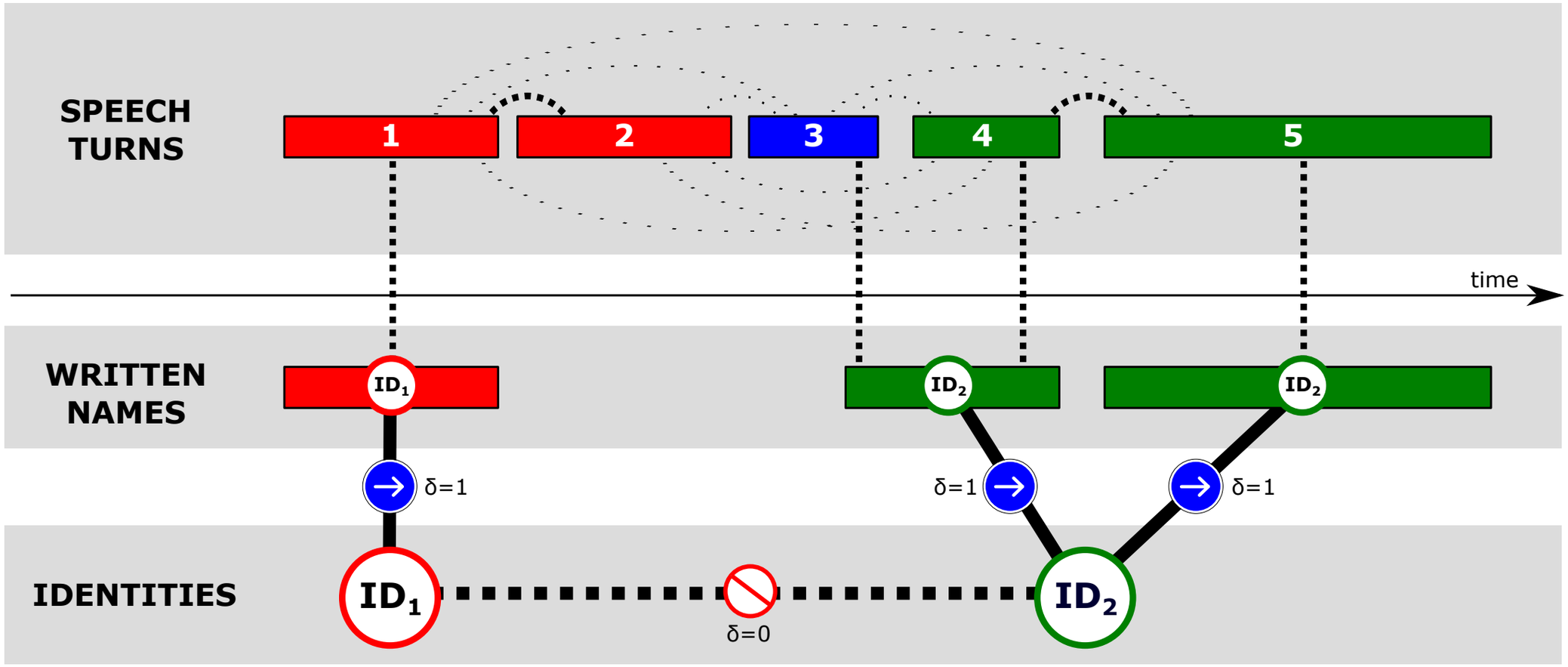
mining

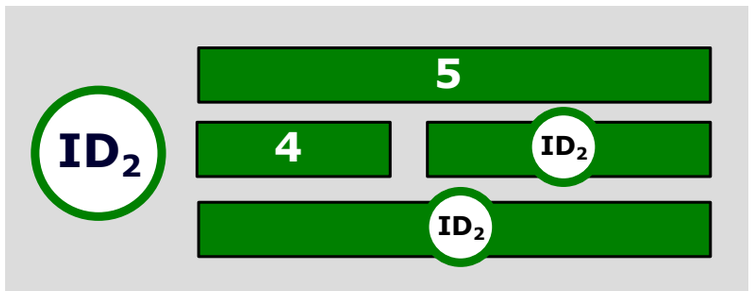
person instance graphs

expected results

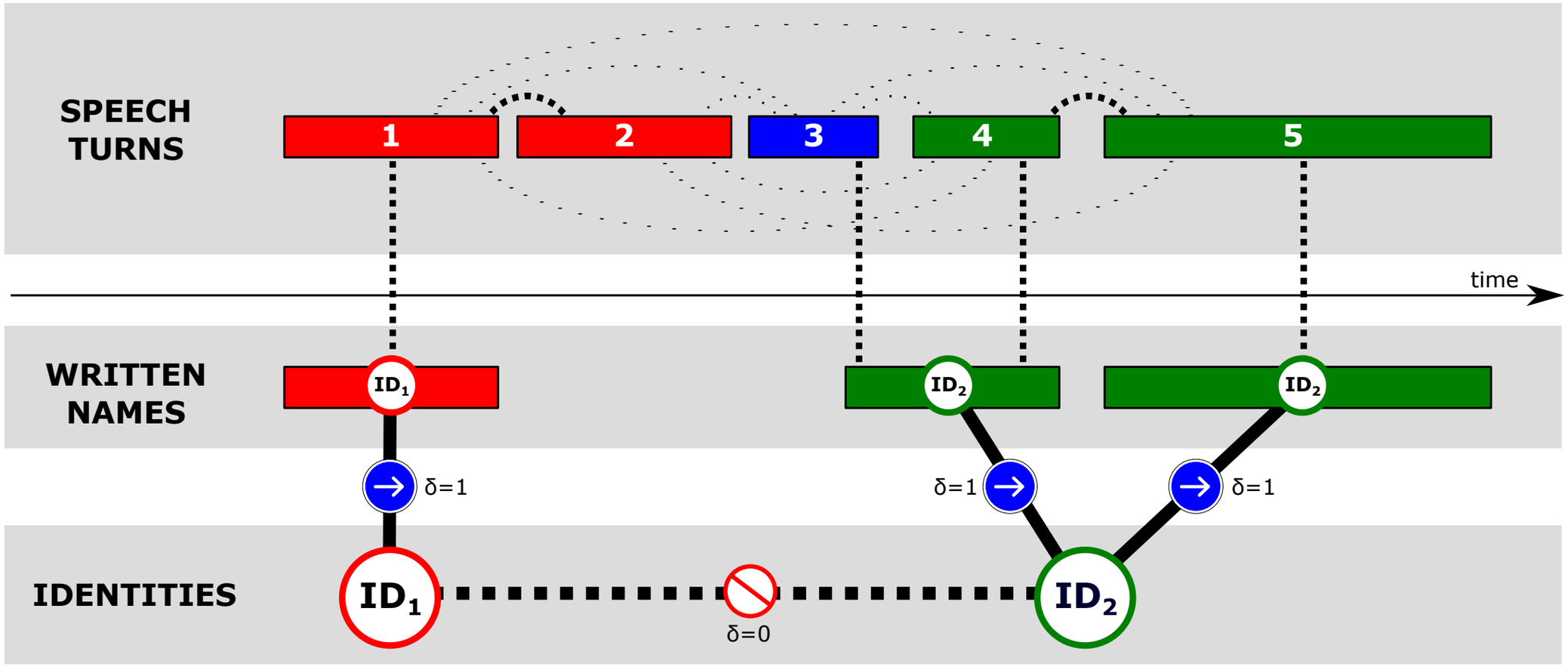


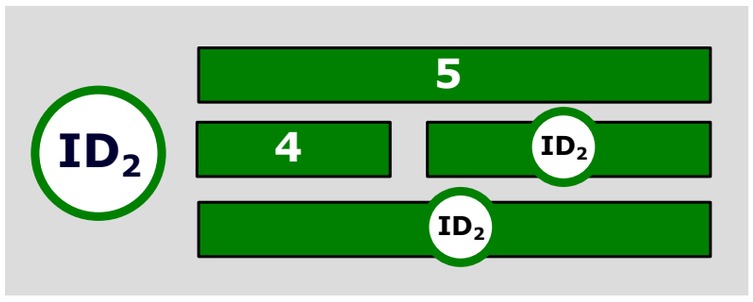
expected results

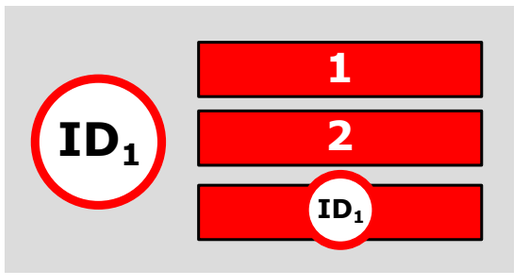




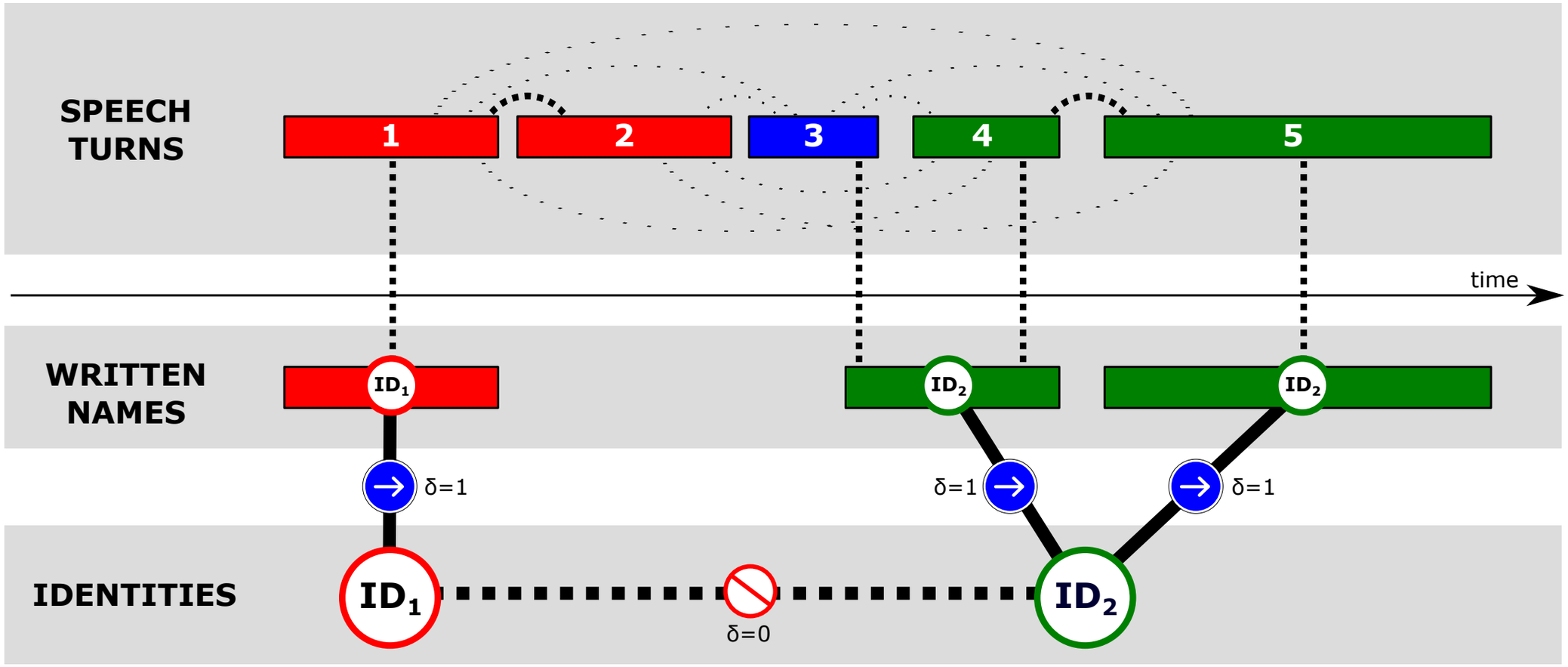
expected results

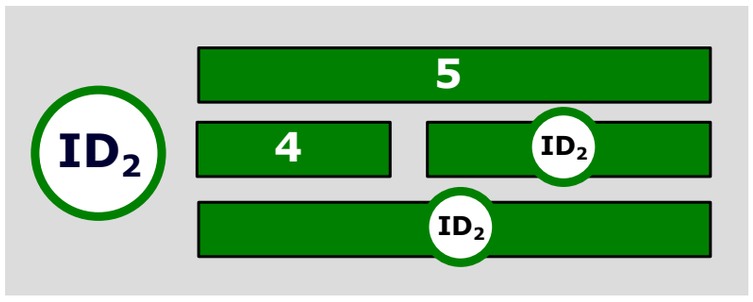


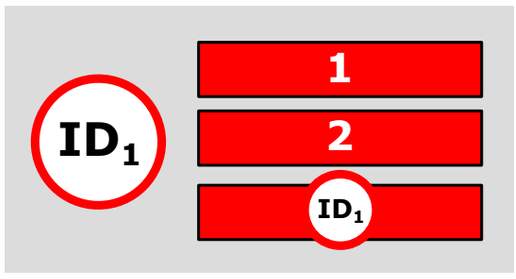


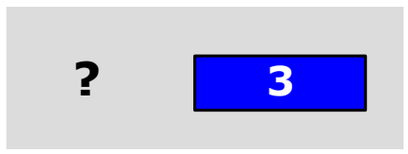


expected results



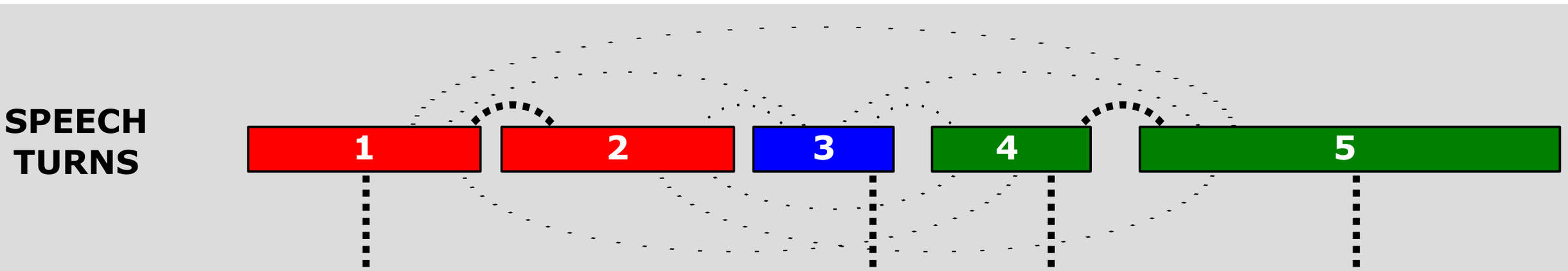






speaker **diarization**

- the task of **partitioning** and **labeling** an audio stream into homogeneous **speech segments** according to the **identity of the speaker**
- usually addressed as a **segmentation** problem followed by a **clustering** problem



three **limitations**

- k-means (et al.) rely on the assumption
  that the **number of clusters** is known ***a priori***.

- hierarchical agglomerative clustering
  does not guarantee **global optimality**
  does not provide clear **stopping criterion**

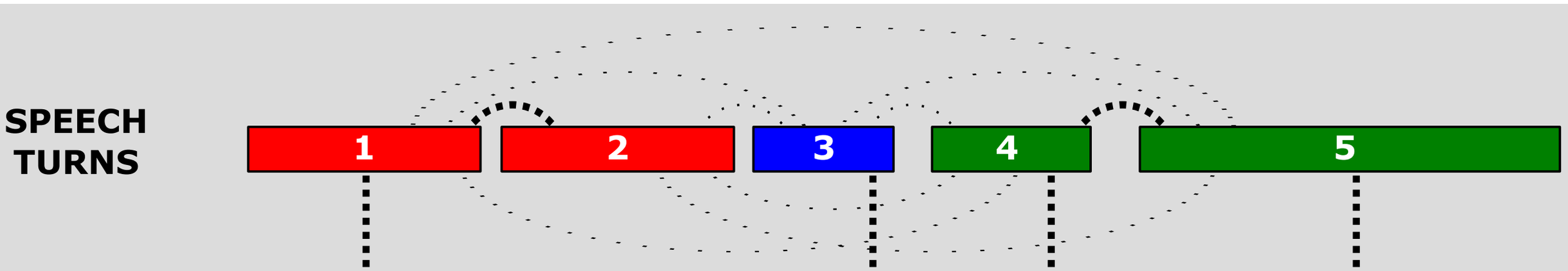
- cannot deal with **incomplete similarity matrices**

- *Finkel & Manning* formulate clustering as an
  integer linear programming problem
  (for co-reference resolution)

notations

- $\mathcal{G} = (\mathcal{V}, \mathcal{E}, p)$ is an **undirected weighted** graph
- $\mathcal{V}$ is the set of **vertices** (**one per item** to cluster)
- $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of **edges**
  two vertices **may or may not be connected**
- edges are weighted by the **probability** that connected vertices belong to the **same cluster**

$$p \colon \mathcal{E} \to [0, 1]$$
$$(v, v') \mapsto p_{vv'} = p(\texttt{ID}(v) = \texttt{ID}(v') \mid v, v')$$



clustering function

- Any output of a valid clustering algorithm can be described by a clustering function

$$\boldsymbol{\delta}\colon \mathcal{V} \times \mathcal{V} \to \{0, 1\}$$

$$(v, v') \mapsto \begin{cases} 1 & \text{if } v \text{ and } v' \text{ are in the same cluster,} \\ 0 & \text{otherwise.} \end{cases}$$

clustering function

- Any output of a valid clustering algorithm can be described by a clustering function

$$\boldsymbol{\delta}\colon \mathcal{V} \times \mathcal{V} \to \{0, 1\}$$

$$(v, v') \mapsto \begin{cases} 1 & \text{if } v \text{ and } v' \text{ are in the same cluster,} \\ 0 & \text{otherwise.} \end{cases}$$

- Reciprocally, additional constraints are needed in order to guarantee a valid clustering

$$\boldsymbol{\delta} \in \Delta_{\mathcal{V}} \qquad \Delta_{\mathcal{V}} = \begin{cases} \boldsymbol{\delta} \in \{0,1\}^{\mathcal{V}\times\mathcal{V}} \text{ s.t. } \forall\, (v, v', v'') \in \mathcal{V}^3, & \\ \text{(a) } \delta_{vv} = 1 & \text{reflexivity} \\ \text{(b) } \delta_{vv'} = \delta_{v'v} & \text{symmetry} \\ \text{(c) } \delta_{vv'} = 1 \wedge \delta_{v'v''} = 1 \implies \delta_{vv''} = 1 & \text{transitivity} \end{cases}$$

objective function

- Find the **clustering function** that maximizes the objective function: $\boldsymbol{\delta}^* = \underset{\boldsymbol{\delta} \in \Delta_{\mathcal{V}}}{\operatorname{argmax}} \ \mathcal{L}^{\alpha}(\boldsymbol{\delta}, \mathcal{E}, p)$

objective function

- Find the **clustering function** that maximizes the objective function: $\boldsymbol{\delta}^* = \underset{\boldsymbol{\delta} \in \Delta_{\mathcal{V}}}{\operatorname{argmax}} \; \mathcal{L}^{\alpha}(\boldsymbol{\delta}, \mathcal{E}, p)$

$$\mathcal{L}^{\alpha}(\boldsymbol{\delta}, \mathcal{E}, p) = |\mathcal{E}|^{-1} \Bigg[\alpha \cdot \overbrace{\sum_{(v,v') \in \mathcal{E}} \delta_{vv'} \cdot p_{vv'}}^{\substack{\text{intra-cluster} \\ \text{similarity}}} + (1-\alpha) \cdot \underbrace{\sum_{(v,v') \in \mathcal{E}} (1-\delta_{vv'}) \cdot (1-p_{vv'})}_{\substack{\text{inter-cluster} \\ \text{dissimilarity}}} \Bigg]$$

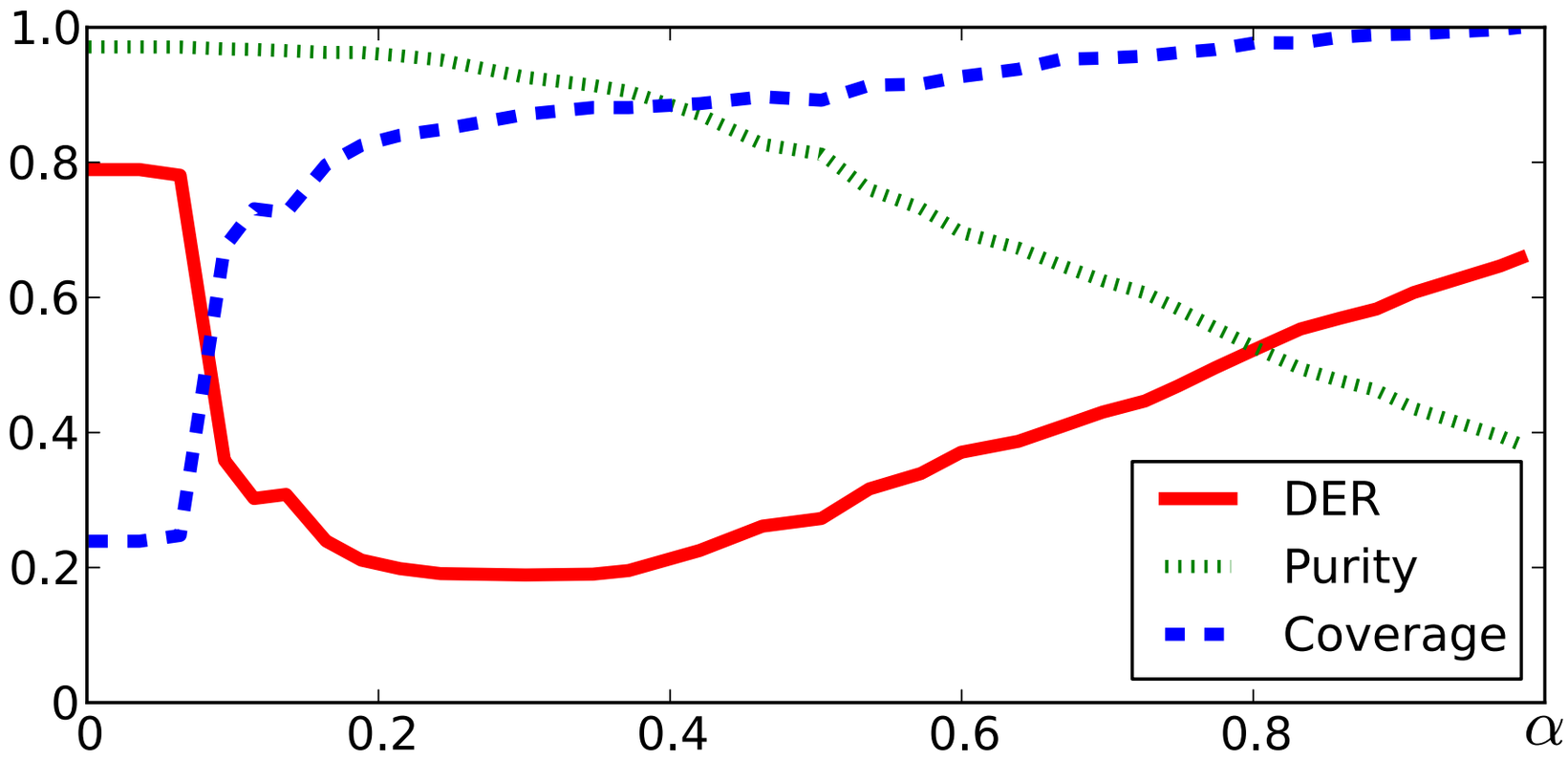
effect of α

$$\mathcal{L}^{\alpha}(\boldsymbol{\delta}, \mathcal{E}, p) = |\mathcal{E}|^{-1}\Big[\alpha \cdot \overbrace{\sum_{(v,v')\in\mathcal{E}} \delta_{vv'} \cdot p_{vv'}}^{\substack{\text{intra-cluster} \\ \text{similarity}}} + (1-\alpha) \cdot \underbrace{\sum_{(v,v')\in\mathcal{E}} (1-\delta_{vv'}) \cdot (1-p_{vv'})}_{\substack{\text{inter-cluster} \\ \text{dissimilarity}}}\Big]$$

effect of **α**

$$\mathcal{L}^{\alpha}(\boldsymbol{\delta}, \mathcal{E}, p) = |\mathcal{E}|^{-1}\Big[\alpha \cdot \overbrace{\sum_{(v,v')\in\mathcal{E}} \delta_{vv'} \cdot p_{vv'}}^{\substack{\text{intra-cluster}\\ \text{similarity}}} + (1-\alpha) \cdot \underbrace{\sum_{(v,v')\in\mathcal{E}} (1-\delta_{vv'}) \cdot (1-p_{vv'})}_{\substack{\text{inter-cluster}\\ \text{dissimilarity}}}\Big]$$

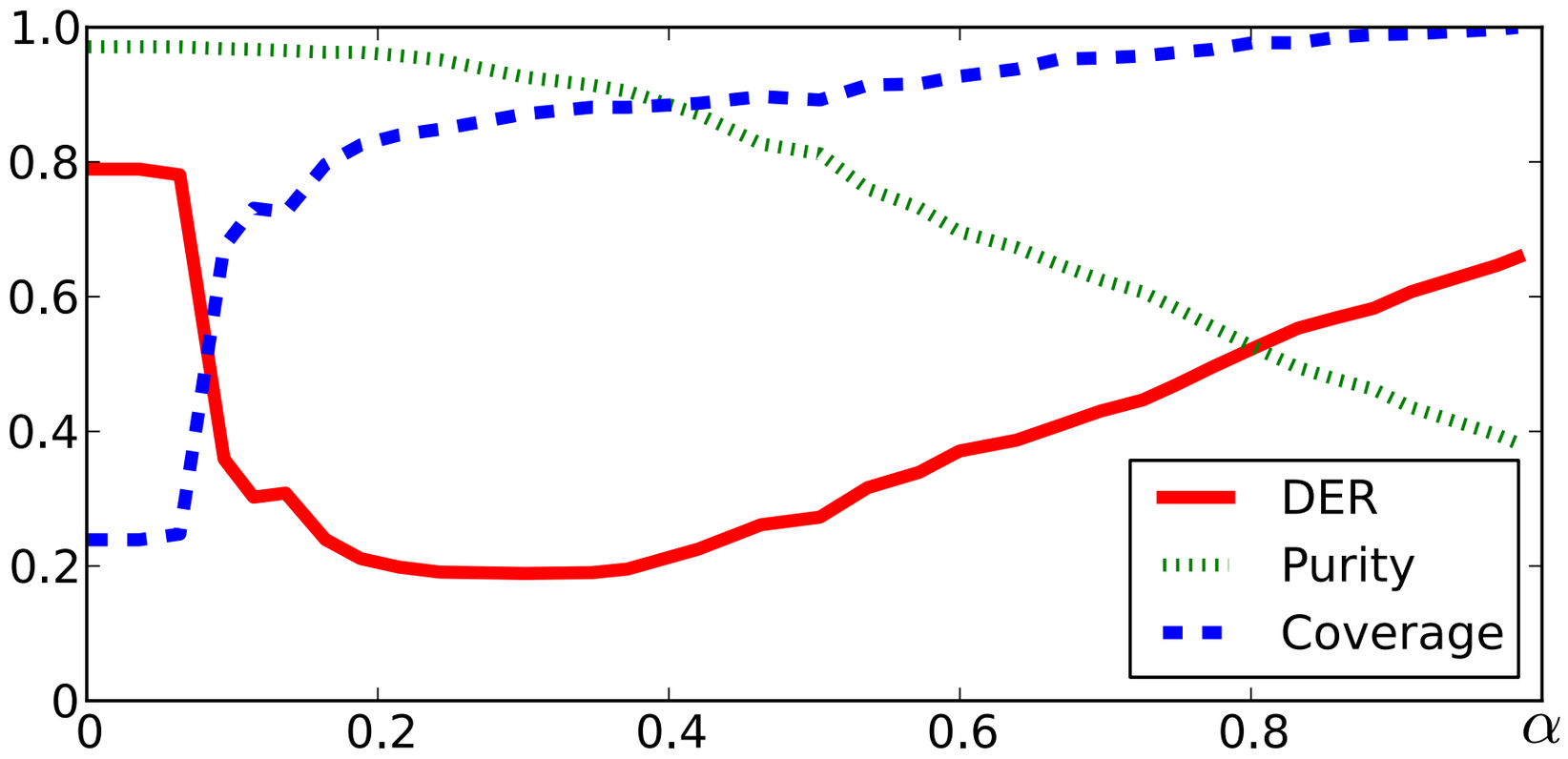
- cluster **purity** *vs.* cluster **coverage**



effect of α

$$\mathcal{L}^{\alpha}(\boldsymbol{\delta}, \mathcal{E}, p) = |\mathcal{E}|^{-1}\Big[\alpha \cdot \overbrace{\sum_{(v,v')\in\mathcal{E}} \delta_{vv'} \cdot p_{vv'}}^{\substack{\text{intra-cluster}\\ \text{similarity}}} + (1-\alpha) \cdot \underbrace{\sum_{(v,v')\in\mathcal{E}} (1-\delta_{vv'}) \cdot (1-p_{vv'})}_{\substack{\text{inter-cluster}\\ \text{dissimilarity}}}\Big]$$

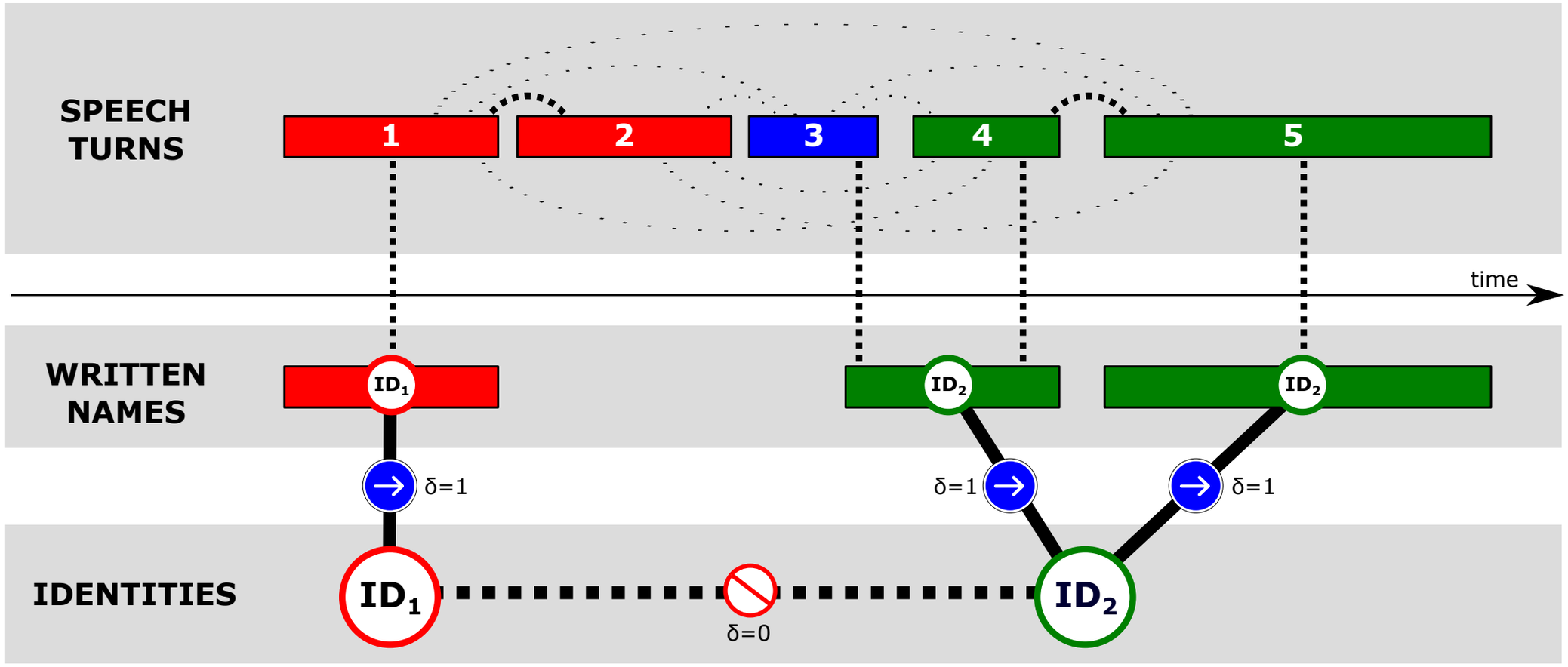
- cluster **purity** *vs.* cluster **coverage**



not nearly as good as standard BIC clustering

😕

extension to **multiple modalities**



multimodal objective function

$$\mathcal{L}^{\alpha}(\boldsymbol{\delta}, \mathcal{E}, p) = |\mathcal{E}|^{-1}\Big[\alpha \cdot \overbrace{\sum_{(v,v')\in\mathcal{E}} \delta_{vv'} \cdot p_{vv'}}^{\substack{\text{intra-cluster} \\ \text{similarity}}} + (1-\alpha) \cdot \underbrace{\sum_{(v,v')\in\mathcal{E}} (1-\delta_{vv'}) \cdot (1-p_{vv'})}_{\substack{\text{inter-cluster} \\ \text{dissimilarity}}}\Big]$$

would result in giving **more weight** to **dominant type** of edges

multimodal objective function

$$\mathcal{L}^{\alpha}(\boldsymbol{\delta}, \mathcal{E}, p) = |\mathcal{E}|^{-1}\Big[\alpha \cdot \overbrace{\sum_{(v,v')\in\mathcal{E}} \delta_{vv'} \cdot p_{vv'}}^{\substack{\text{intra-cluster}\\ \text{similarity}}} + (1-\alpha) \cdot \underbrace{\sum_{(v,v')\in\mathcal{E}} (1-\delta_{vv'}) \cdot (1-p_{vv'})}_{\substack{\text{inter-cluster}\\ \text{dissimilarity}}}\Big]$$

would result in giving **more weight** to **dominant type** of edges

$$\mathcal{L}^{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}(\boldsymbol{\delta}, \mathcal{E}, p) = \sum_{\substack{x\in\{\mathcal{T},\mathcal{W},\mathcal{S},\mathcal{I}\}\\ y\in\{\mathcal{T},\mathcal{W},\mathcal{S},\mathcal{I}\}}} \beta_{xy} \cdot \mathcal{L}^{\alpha_{xy}}\left(\boldsymbol{\delta}, \mathcal{E} \cap (x \times y), p\right) \qquad \alpha_{xy} \in [0,1] \quad \beta_{xy} \in [0,1] \quad \sum_{x,y} \beta_{xy} = 1$$

one objective function per edge type subgraph
allows different weights beta for different edge types
allows fine tuning of alpha for each edge type

multimodal objective function

$$\mathcal{L}^{\alpha}(\boldsymbol{\delta}, \mathcal{E}, p) = |\mathcal{E}|^{-1}\Big[\alpha \cdot \overbrace{\sum_{(v,v')\in\mathcal{E}} \delta_{vv'} \cdot p_{vv'}}^{\substack{\text{intra-cluster}\\ \text{similarity}}} + (1-\alpha) \cdot \underbrace{\sum_{(v,v')\in\mathcal{E}} (1-\delta_{vv'}) \cdot (1-p_{vv'})}_{\substack{\text{inter-cluster}\\ \text{dissimilarity}}}\Big]$$

would result in giving **more weight** to **dominant type** of edges

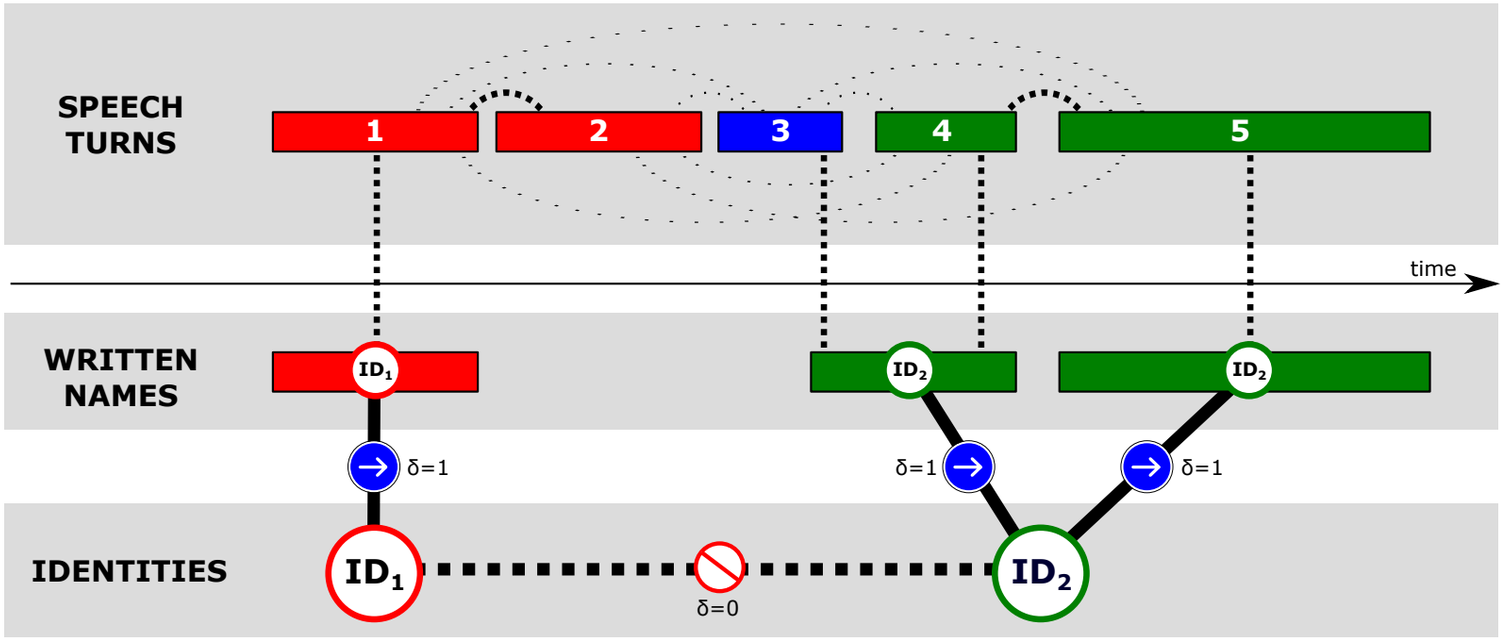
$$\mathcal{L}^{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}(\boldsymbol{\delta}, \mathcal{E}, p) = \sum_{\substack{x\in\{\mathcal{T},\mathcal{W},\mathcal{S},\mathcal{I}\}\\ y\in\{\mathcal{T},\mathcal{W},\mathcal{S},\mathcal{I}\}}} \beta_{xy} \cdot \mathcal{L}^{\alpha_{xy}}\left(\boldsymbol{\delta}, \mathcal{E}\cap(x\times y), p\right) \qquad \alpha_{xy} \in [0,1] \qquad \beta_{xy} \in [0,1] \quad \sum_{x,y}\beta_{xy} = 1$$

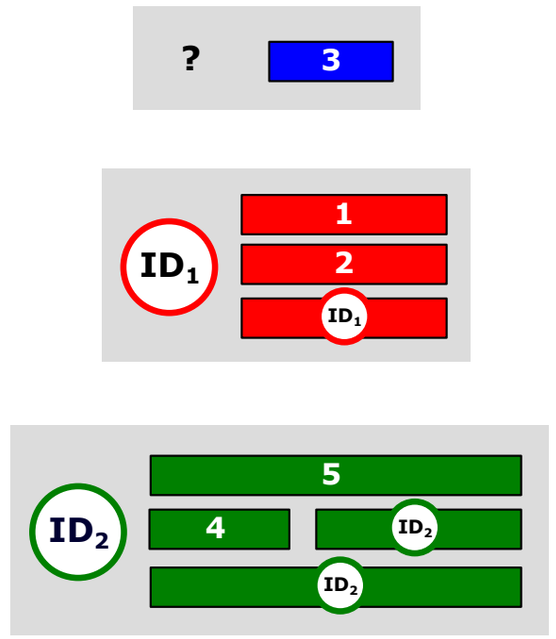
one objective function per edge type subgraph
allows different weights beta for different edge types
allows fine tuning of alpha for each edge type

$$\boldsymbol{\delta}^* = \underset{\boldsymbol{\delta}\in\Delta_{\mathcal{V}}}{\operatorname{argmax}}\ \mathcal{L}^{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}(\boldsymbol{\delta}, \mathcal{E}, p)$$

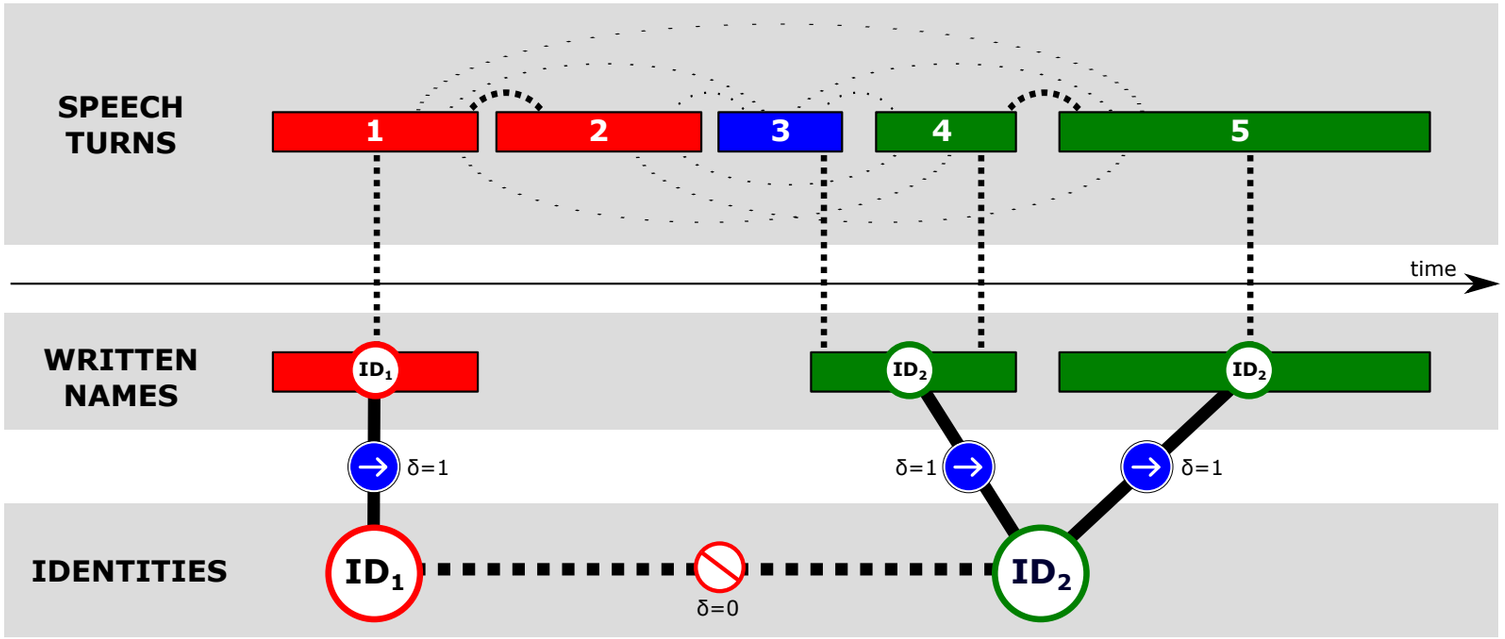
$\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are optimized by random search

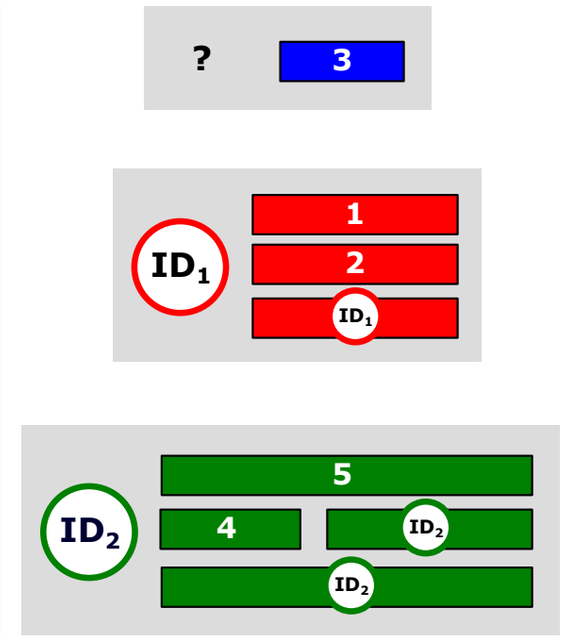
additional **constraints**





additional **constraints**





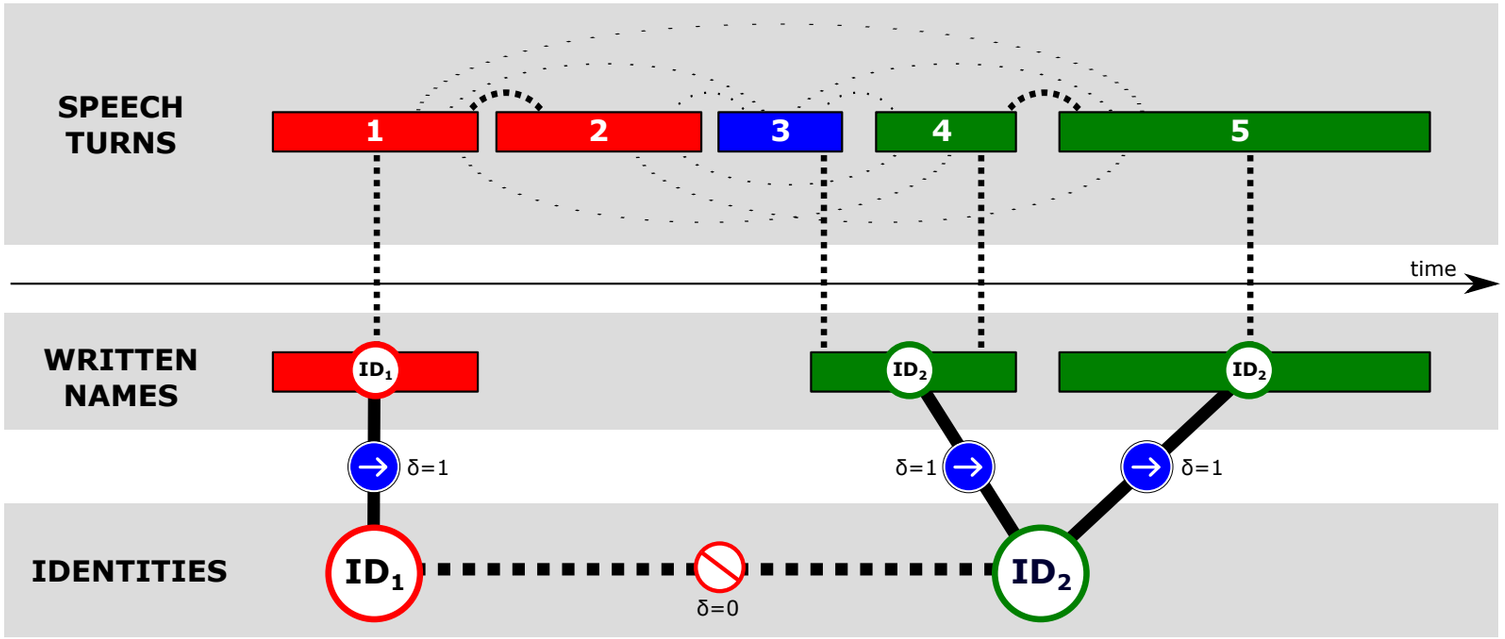
two identity vertices
cannot end up
in the same cluster

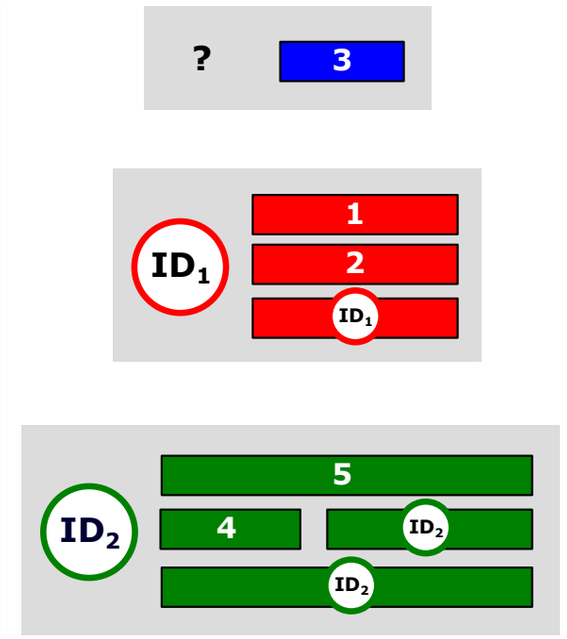
$$\forall\,(i,i') \in \mathcal{I}^2,\ \ i \neq i' \implies \delta_{ii'} = 0$$

any instance has at most one identity

$$\forall v \in \mathcal{V},\ \ \sum_{i \in \mathcal{I}} \delta_{vi} \leq 1$$

additional **constraints**





two identity vertices
cannot end up
in the same cluster

$$\forall\,(i,i') \in \mathcal{I}^2,\ \ i \neq i' \implies \delta_{ii'} = 0$$

any written or spoken instance
must be in the same cluster
as their corresponding identity

$$\forall w \in \mathcal{W},\ \ \delta_{wi_w} = 1$$

$$\forall s \in \mathcal{S},\ \ \delta_{si_s} = 1$$

any instance has at most one identity

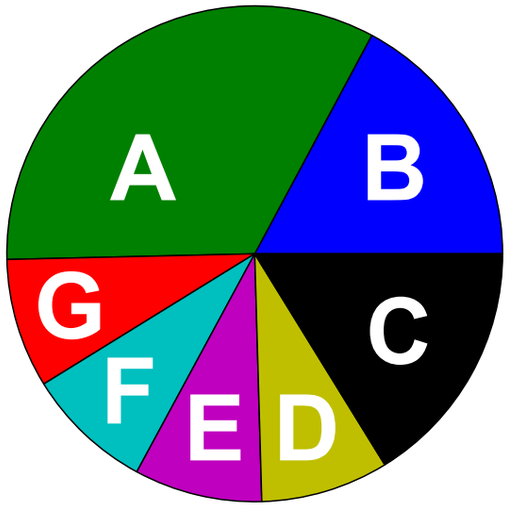
$$\forall v \in \mathcal{V},\ \ \sum_{i \in \mathcal{I}} \delta_{vi} \leq 1$$

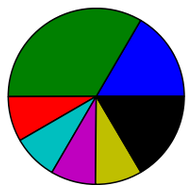
experimental **protocol**

- corpora

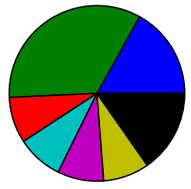
TRAINING SET
24 hours

A: BFM Story
B: LCP Info
C: Top Questions
D: Ça Vous Regarde
E: Planète Showbiz
F: Entre Les Lignes
G: Pile Et Face





DEVELOPMENT
3 hours



TEST
3 hours

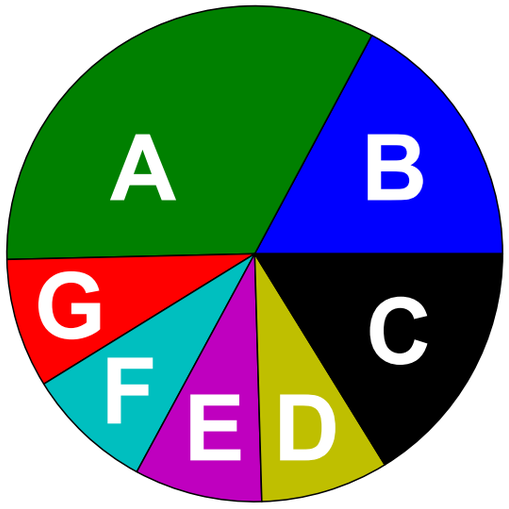
- evaluation metrics
  diarization error rate
  identification error rate / **precision** / **recall**

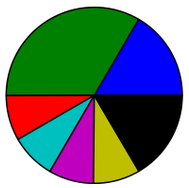
experimental **protocol**

- corpora

TRAINING SET
24 hours

A: BFM Story
B: LCP Info
C: Top Questions
D: Ça Vous Regarde
E: Planète Showbiz
F: Entre Les Lignes
G: Pile Et Face





DEVELOPMENT
3 hours

TEST
3 hours

$$p_{tt} \quad p_{ti} \quad p_{tw} \quad p_{ts}$$

- evaluation metrics
  diarization error rate
  identification error rate / **precision** / **recall**

experimental **protocol**

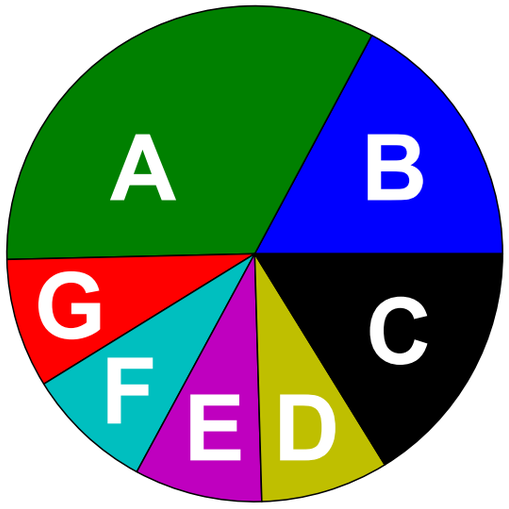
- corpora

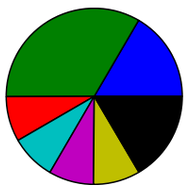
$$\mathrm{XER} \in \{\mathrm{IER}, \mathrm{DER}\}$$

$$(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} \ \mathbb{E}_{\mathrm{dev}} \left[\mathrm{XER}(r, h)\right]$$

TRAINING SET
24 hours

A: BFM Story
B: LCP Info
C: Top Questions
D: Ça Vous Regarde
E: Planète Showbiz
F: Entre Les Lignes
G: Pile Et Face





DEVELOPMENT
3 hours

TEST
3 hours

$$p_{tt} \quad p_{ti} \quad p_{tw} \quad p_{ts}$$

- evaluation metrics
  diarization error rate
  identification error rate / **precision** / **recall**

different graph **configuration** for different **tasks**

- speaker diarization
  speech turn vertices only

- speaker identification

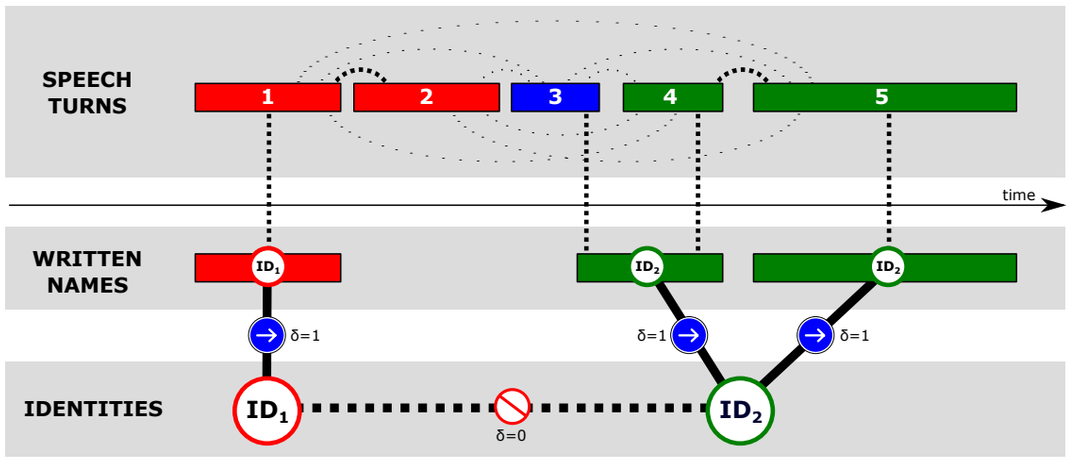
- **supervised mono-modal** speaker identification
  no written or spoken vertices

- **unsupervised cross-modal** speaker identification
  no supervised edges
  only written or spoken vertices

- **multi-modal** speaker identification
  the best of both worlds

results **speaker diarization**

Vertices	**Edges**	**DER**	**Purity**	**Coverage**
$\mathcal{T}$	$t \leftrightarrow t'$	**21.1%**	**94.7%**	83.6%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}}$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	18.3%	93.4%	85.8%
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	18.2%	94.0%	86.1%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	**17.8%**	93.9%	85.7%
BIC clustering baseline [1]		**19.8%**	**92.1%**	86.8%



written names do help 😌

supervision does help 😊

complementarity 😃

results **mono-modal speaker identification**

Vertices	**Edges**	**IER**	**P.**	**R.**
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t$	49.4%	54.7%	54.3%
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	**47.9%**	55.9%	**55.8%**
GMM-UBM baseline		**48.9%**	57.5%	54.3%

- the first configuration is equivalent to usual open-set speaker identification

$$\texttt{ID}(t) = \begin{cases} i^* = \underset{i \in \mathcal{I}^*}{\text{argmax}}\ p_{ti} & \text{if } p_{ti^*} > (1 - \alpha_{\mathcal{T}\mathcal{I}^*}) \\ \textcircled{?} & \text{otherwise.} \end{cases}$$

- GMM-UBM baseline relies on preliminary speaker diarization step and performs identification at cluster level

- adding $t \leftrightarrow t'$ edges addresses this limitation

$\beta_{\mathcal{T}\mathcal{T}} = 0.55$ $\beta_{\mathcal{T}\mathcal{I}} = 0.45$ $\alpha_{\mathcal{T}\mathcal{T}} = 0.19$

results **mono-modal speaker identification**

Vertices	Edges	IER	P.	R.
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t$	49.4%	54.7%	54.3%
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	**47.9%**	55.9%	**55.8%**
GMM-UBM baseline		**48.9%**	57.5%	54.3%

😀

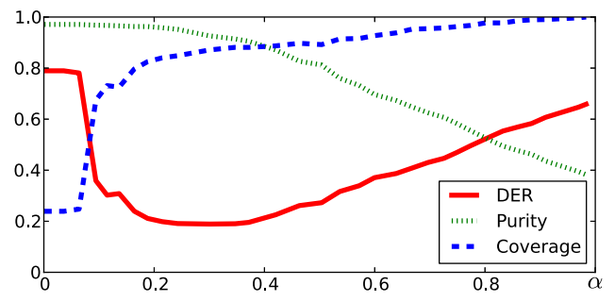
- the first configuration is equivalent to usual open-set speaker identification

$$\texttt{ID}(t) = \begin{cases} i^* = \underset{i \in \mathcal{I}^*}{\text{argmax}}\ p_{ti} & \text{if } p_{ti^*} > (1 - \alpha_{\mathcal{T}\mathcal{I}^*}) \\ \textcircled{?} & \text{otherwise.} \end{cases}$$

- GMM-UBM baseline relies on preliminary speaker diarization step and performs identification at cluster level

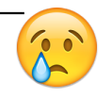
- adding $t \leftrightarrow t'$ edges addresses this limitation

$\beta_{\mathcal{T}\mathcal{T}} = 0.55$ $\beta_{\mathcal{T}\mathcal{I}} = 0.45$ $\alpha_{\mathcal{T}\mathcal{T}} = 0.19$



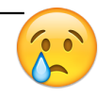
results **cross-modal speaker identification**

Vertices	**Edges**	**IER**	**Precision**	**Recall**
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}}$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	**46.5**%	66.8%	56.9%
$\mathcal{T} \cup \mathcal{S} \cup \mathcal{I}_{\mathcal{S}}$	$t \leftrightarrow t' \leftrightarrow s \Leftrightarrow i_s$	81.8%	21.5%	21.4%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}} \cup \mathcal{S} \cup \mathcal{I}_{\mathcal{S}}$	$i_w \Leftrightarrow w \leftrightarrow t \leftrightarrow t' \leftrightarrow s \Leftrightarrow i_s$	**45.6**%	62.7%	58.2%



results **cross-modal speaker identification**

Vertices	**Edges**	**IER**	**Precision**	**Recall**
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}}$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	**46.5**%	66.8%	56.9%
$\mathcal{T} \cup \mathcal{S} \cup \mathcal{I}_{\mathcal{S}}$	$t \leftrightarrow t' \leftrightarrow s \Leftrightarrow i_s$	81.8%	21.5%	21.4%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}} \cup \mathcal{S} \cup \mathcal{I}_{\mathcal{S}}$	$i_w \Leftrightarrow w \leftrightarrow t \leftrightarrow t' \leftrightarrow s \Leftrightarrow i_s$	**45.6**%	62.7%	58.2%



- to be compared with *oracle* performance

results **cross-modal speaker identification**

Vertices	**Edges**	**IER**	**Precision**	**Recall**
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_\mathcal{W}$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	**46.5%**	66.8%	56.9%
$\mathcal{T} \cup \mathcal{S} \cup \mathcal{I}_\mathcal{S}$	$t \leftrightarrow t' \leftrightarrow s \Leftrightarrow i_s$	81.8%	21.5%	21.4%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_\mathcal{W} \cup \mathcal{S} \cup \mathcal{I}_\mathcal{S}$	$i_w \Leftrightarrow w \leftrightarrow t \leftrightarrow t' \leftrightarrow s \Leftrightarrow i_s$	**45.6%**	62.7%	58.2%

- to be compared with *oracle* performance

Vertices	**IER**	**Precision**	**Recall**
$\mathcal{I}_\mathcal{W}$	**39.3%**	100.0%	**63.8%**
$\mathcal{I}_\mathcal{S}$	35.8%	100.0%	67.3%
$\mathcal{I}_\mathcal{W} \cup \mathcal{I}_\mathcal{S}$	21.5%	100.0%	81.9%

results **multi-modal speaker identification**

Vertices	Edges	All			Anchors			All but anchors		
		IER	**P.**	**R.**	**IER**	**P.**	**R.**	**IER**	**P.**	**R.**
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	47.9%	55.9%	55.8%	**20.3**%	**86.6**%	**79.7**%	51.8%	50.8%	48.3%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}}$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	46.5%	66.8%	56.8%	79.4%	31.8%	20.6%	34.4%	76.9%	65.8%
$\mathcal{T} \cup \mathcal{I}^* \cup \mathcal{W} \cup \mathcal{I}_{\mathcal{W}}$	$i^* \leftrightarrow t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	**25.3**%	**79.4**%	**78.6**%	23.9%	82.9%	76.1%	**22.4**%	**82.5**%	**77.9**%

results **multi-modal speaker identification**

Vertices	Edges	All			Anchors			All but anchors		
		IER	**P.**	**R.**	**IER**	**P.**	**R.**	**IER**	**P.**	**R.**
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	47.9%	55.9%	55.8%	**20.3**%	**86.6**%	**79.7**%	51.8%	50.8%	48.3%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_\mathcal{W}$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	46.5%	66.8%	56.8%	79.4%	31.8%	20.6%	34.4%	76.9%	65.8%
$\mathcal{T} \cup \mathcal{I}^* \cup \mathcal{W} \cup \mathcal{I}_\mathcal{W}$	$i^* \leftrightarrow t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	**25.3**%	**79.4**%	**78.6**%	23.9%	82.9%	76.1%	**22.4**%	**82.5**%	**77.9**%

astonished face 

results **multi-modal speaker identification**

Vertices	Edges	All			Anchors			All but anchors		
		IER	**P.**	**R.**	**IER**	**P.**	**R.**	**IER**	**P.**	**R.**
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	47.9%	55.9%	55.8%	**20.3**%	**86.6**%	**79.7**%	51.8%	50.8%	48.3%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_\mathcal{W}$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	46.5%	66.8%	56.8%	79.4%	31.8%	20.6%	34.4%	76.9%	65.8%
$\mathcal{T} \cup \mathcal{I}^* \cup \mathcal{W} \cup \mathcal{I}_\mathcal{W}$	$i^* \leftrightarrow t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	**25.3**%	**79.4**%	**78.6**%	23.9%	82.9%	76.1%	**22.4**%	**82.5**%	**77.9**%

astonished face 

 grinning face

results **multi-modal speaker identification**

Vertices	Edges	All			Anchors			All but anchors		
		IER	**P.**	**R.**	**IER**	**P.**	**R.**	**IER**	**P.**	**R.**
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	47.9%	55.9%	55.8%	**20.3**%	**86.6**%	**79.7**%	51.8%	50.8%	48.3%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_\mathcal{W}$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	46.5%	66.8%	56.8%	79.4%	31.8%	20.6%	34.4%	76.9%	65.8%
$\mathcal{T} \cup \mathcal{I}^* \cup \mathcal{W} \cup \mathcal{I}_\mathcal{W}$	$i^* \leftrightarrow t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	**25.3**%	**79.4**%	**78.6**%	23.9%	82.9%	76.1%	**22.4**%	**82.5**%	**77.9**%

astonished face 

 grinning face

- **the best of both worlds**
  - supervised for anchors
    lots of training data
  - cross-modal for guests or reporters
    often introduced by overlaid names
- **com-ple-men-ta-ry** approaches

results **multi-modal speaker identification**

Vertices	Edges	All			Anchors			All but anchors		
		IER	**P.**	**R.**	**IER**	**P.**	**R.**	**IER**	**P.**	**R.**
$\mathcal{T} \cup \mathcal{I}^*$	$i^* \leftrightarrow t \leftrightarrow t'$	47.9%	55.9%	55.8%	**20.3**%	**86.6**%	**79.7**%	51.8%	50.8%	48.3%
$\mathcal{T} \cup \mathcal{W} \cup \mathcal{I}_\mathcal{W}$	$t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	46.5%	66.8%	56.8%	79.4%	31.8%	20.6%	34.4%	76.9%	65.8%
$\mathcal{T} \cup \mathcal{I}^* \cup \mathcal{W} \cup \mathcal{I}_\mathcal{W}$	$i^* \leftrightarrow t \leftrightarrow t' \leftrightarrow w \Leftrightarrow i_w$	**25.3**%	**79.4**%	**78.6**%	23.9%	82.9%	76.1%	**22.4**%	**82.5**%	**77.9**%

- **the best of both worlds**
  - supervised for anchors
    lots of training data
  - cross-modal for guests or reporters
    often introduced by overlaid names
- **com-ple-men-ta-ry** approaches

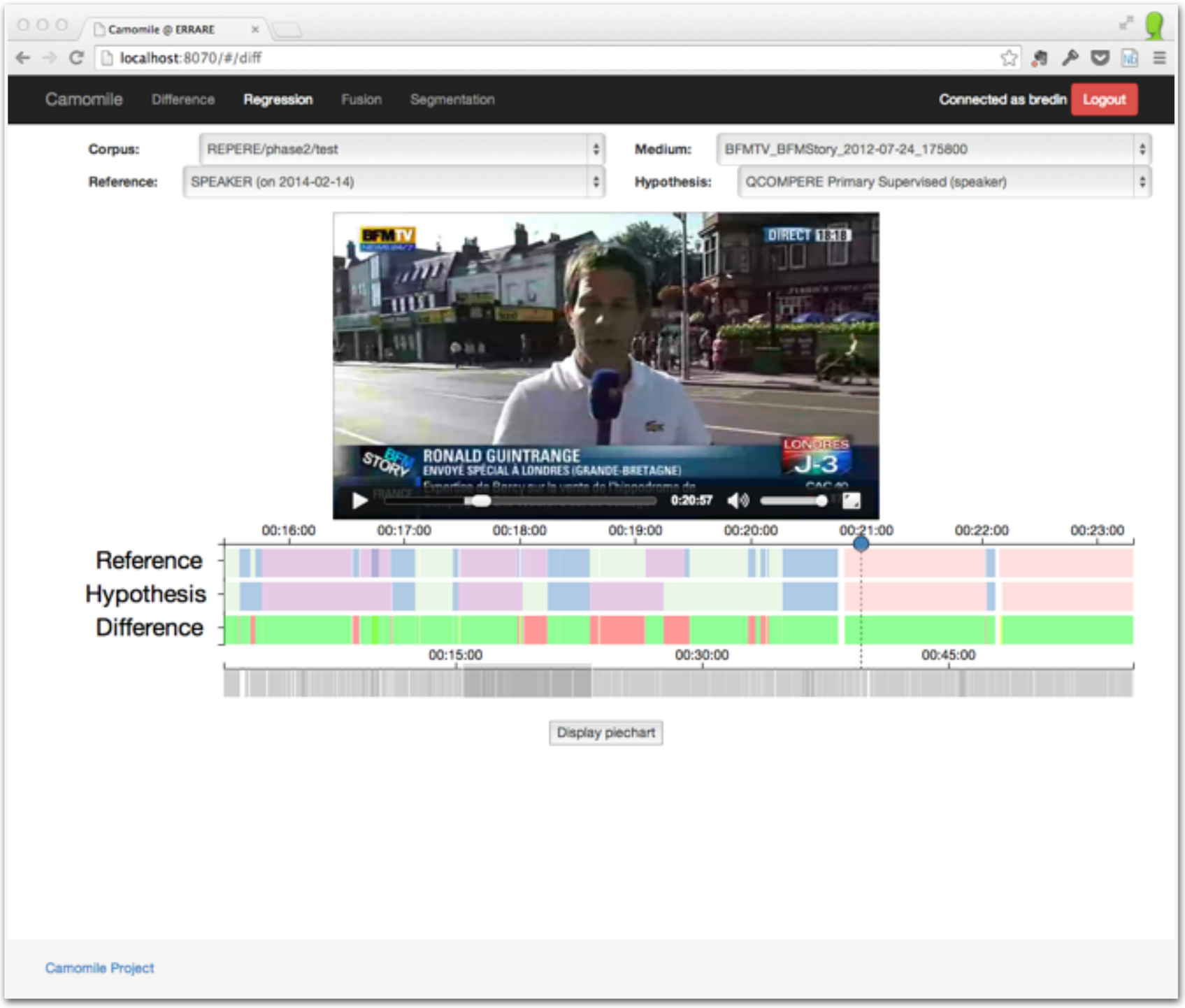
astonished face

grinning face

confused face

some room for improvement about anchors

dem


Camomile @ ERRARE
localhost:8070/#/diff
Camomile
Difference
Regression
Fusion
Segmentation
Connected as bredin
Logout
Corpus:
REPERE/phase2/test
Medium:
BFMTV_BFMStory_2012-07-24_175800
Reference:
SPEAKER (on 2014-02-14)
Hypothesis:
QCOMPERE Primary Supervised (speaker)
BFMTV
DIRECT 18:18
RONALD GUINTRANGE
ENVOYÉ SPÉCIAL À LONDRES (GRANDE-BRETAGNE)
LONDRES
J-3
0:20:57
00:16:00
00:17:00
00:18:00
00:19:00
00:20:00
00:21:00
00:22:00
00:23:00
Reference
Hypothesis
Difference
00:15:00
00:30:00
00:45:00
Display piechart
Camomile Project

conclusion

- a global **optimisation framework**
  for **person identification** in multimedia documents

- a **unified graphical** approach
  for mono-, cross- and multi-modal identification
  easily extendable to additional modalities

- **main limitations — scalability**
  problem quickly becomes intractable
  as the number of vertices increases

what's **next**?

- extension to **face** modality

  $f \leftrightarrow f'$ face clustering

  $f \leftrightarrow i$ supervised face recognition

  $t \leftrightarrow f$ audio-based face recognition

- **alternative** objective functions

  integrate **prior knowledge** about the temporal **structure** of interactions

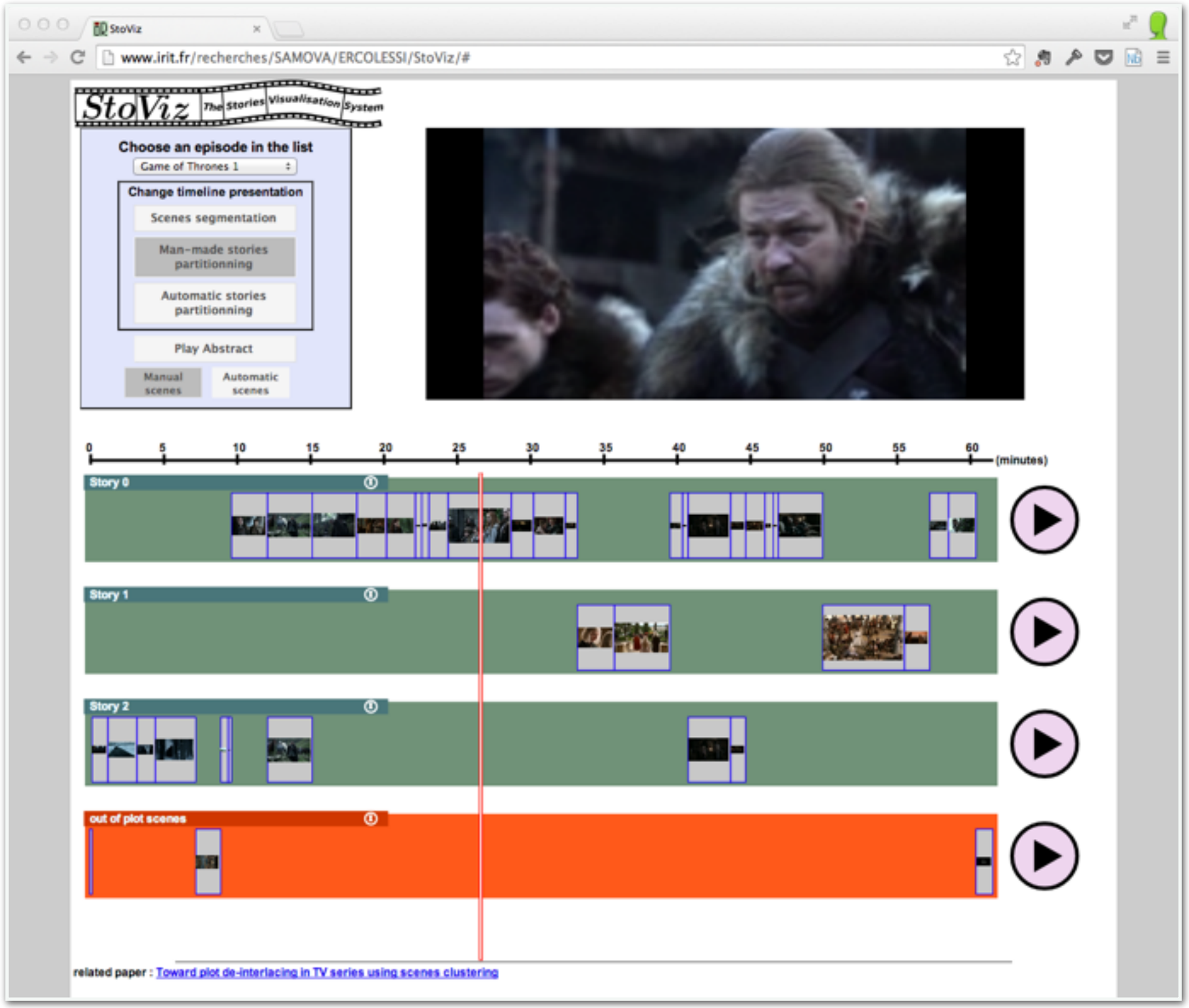
- **named speaker identification**

  better use of spoken names when they are the only source of information

- other **multimodal clustering** problems

  scene clustering for TV series plot de-interlacing

stoviz


StoViz
www.irit.fr/recherches/SAMOVA/ERCOLESSI/StoViz/#
StoViz The Stories Visualisation System
Choose an episode in the list
Game of Thrones 1
Change timeline presentation
Scenes segmentation
Man-made stories partitionning
Automatic stories partitionning
Play Abstract
Manual scenes
Automatic scenes
0 5 10 15 20 25 30 35 40 45 50 55 60 (minutes)
Story 0
Story 1
Story 2
out of plot scenes
related paper : Toward plot de-interlacing in TV series using scenes clustering

person instance **graphs** for person **recognition** in **multimedia** data

Hervé Bredin, Anindya Roy, Viet-Bac Le and Claude Barras
Person Instance Graphs for Mono-, Cross- and Multi-Modal Person Recognition in Multimedia Data: Application to Speaker Identification in TV Broadcast
International Journal of Multimedia Information Retrieval, May 2014

Hervé Bredin, Antoine Laurent, Achintya Sarkar, Viet-Bac Le, Claude Barras and Sophie Rosset
Person Instance Graphs for Named Speaker Identification
Odyssey 2014: the Speaker and Language Recognition Workshop, June 2014

Anindya Roy, Camille Guinaudeau, Hervé Bredin and Claude Barras
TVD: a Reproducible and Multiply Aligned TV Series Dataset
LREC 2014: the 9th edition of the Language Resources and Evaluation Conference, May 2014

tvd.niderb.fr