GRAMS: A Graph-based Approach for Inferring Semantic Descriptions of Wikipedia Tables

Binh Vu, Craig Knoblock, Pedro Szekely, Jay Pujara, Minh Pham

Information Sciences Institute
University of Southern California

Motivating Example

- Wikipedia has 7.5 millions tables covering many domains

List of albums in 2019 (USA)

Date	Album	Artist	Genre (s)
5	*Metawar*	3Teeth	Industrial · industrial metal
	Stonechild	Jesca Hoop	Folk · blues · pop
	Hotel Diablo	Machine Gun Kelly	Hip hop
	FLYGOD is an Awesome GOD	Westside Gunn	Hip hop

Motivating Example

- Wikipedia has 7.5 millions tables covering many domains

List of public schools in New South Wales

Name	Suburb	LGA	Opened	Website
Adaminaby Public School	Adaminaby	Snowy Monaro	1869	Website
Albion Park Public School	Albion Park	Shellharbour	1872	Website
Albion Park Rail Public School	Albion Park Rail	Shellharbour	1959	Website

List of albums in 2019 (USA)

Date	Album		
5	*Metawar*		
	Stonechild		
	Hotel Diablo		
	FLYGOD is an Awesome GOD	Westside Gunn	Hip hop

Motivating Example

- Wikipedia has 7.5 millions tables covering many domains

List of historic railway stations

Name on closure	Place	Opened	Closed to passengers	Railway company
Abbey & West Dereham	West Dereham	1882	1930	Great Eastern
Aldeby	Aldeby	1854	1959	Great Eastern
Ashwellthorpe	Ashwellthorpe	1881	1939	Great Eastern

ic schools in New South Wales

		Opened	Website
	haro	1869	Website
	ur	1872	Website
School	Rail	Shellharbour	1959	Website

Hotel Diablo			
FLYGOD is an Awesome GOD		Westside Gunn	Hip hop

Motivating Example

- Wikipedia has 7.5 millions tables covering many domains

List of drugs granted breakthrough therapy designation

Drug	Manufacturer	Indication
Psilocybin	Usona Institute	major depressive disorder[1]
B38M (JNJ-4528)	Legend Biotech/Janssen	multiple myeloma
Rilonacept	Kiniska Pharmaceuticals	recurrent pericarditis
Inmazeb	Regeneron	Ebola Virus
Olorofim	F2G	invasive mold infections

List of historic railway stati

Name on closure	Place			
Abbey & West Dereham	West Dereham			
Aldeby	Aldeby			
Ashwellthorpe	Ashwellthorpe	1881	1939	Great Eastern

Hotel Diablo	School	Rail	Shellharbour	1959	Website
FLYGOD is an Awesome GOD		Westside Gunn	Hip hop		

Motivating Example

- Wikipedia has 7.5 millions tables covering many domains

Members of 56th New Brunswick Legislature

Name	Party	Riding
Hédard Albert	Liberal	Caraquet
David Alward	Progressive Conservative	Woodstock
Donald Arseneault	Liberal	Dalhousie-Restigouche East
John Betts	Progressive Conservative	Moncton Crescent

kthrough therapy designation

	Indication
	depressive disorder[1]
	e myeloma
	nt pericarditis
Inmazeb	Regeneron	Ebola Virus
Olorofim	F2G	invasive mold infections

Dereham	Dereham				
Aldeby	Aldeby				
Ashwellthorpe	Ashwellthorpe	1881	1939	Great Eastern	
Hotel Diablo		School	Rail	Shellharbour	1959	Website
FLYGOD is an Awesome GOD			Westside Gunn	Hip hop	

Motivating Example

- Wikipedia has 7.5 millions tables covering many domains

List of players won Walter Payton Award

Year	Player	School	Position
1987	Kenny Gamble	Colgate	RB
1988	Dave Meggett	Towson State	RB
2018	Devlin Hodges	Samford	QB
2019	Trey Lance	North Dakota State	QB

Members of 56th New Brunswick Legislature

Name	Party	Ri
Hédard Albert	Liberal	Caraquet
David Alward	Progressive Conservative	Woodstock
Donald Arseneault	Liberal	Dalhousie-Restigouche East
John Betts	Progressive Conservative	Moncton Crescent

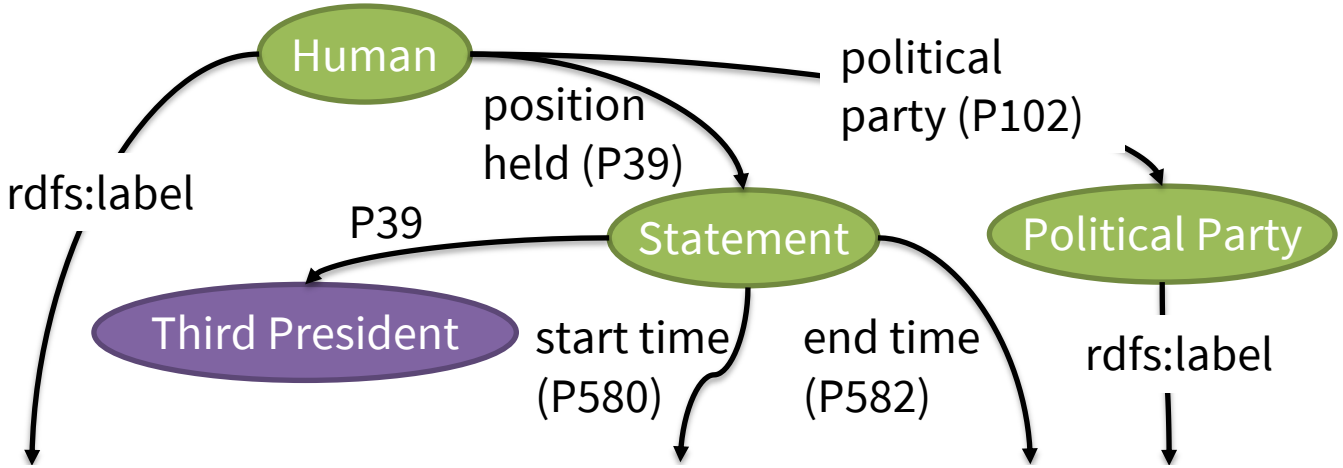
Dereham	Dereham	Inmazeb	Regeneron	Ebola Virus
Aldeby	Aldeby	Olorofim	F2G	invasive mold infections
Ashwellthorpe	Ashwellthorpe	1881	1939	Great Eastern

Hotel Diablo	School	Rail	Shellharbour	1959	Website
FLYGOD is an Awesome GOD		Westside Gunn	Hip hop		

nt pericarditis

Source Modeling Problem

- Building semantic descriptions of tables
  - Describing data source using classes and properties in ontologies

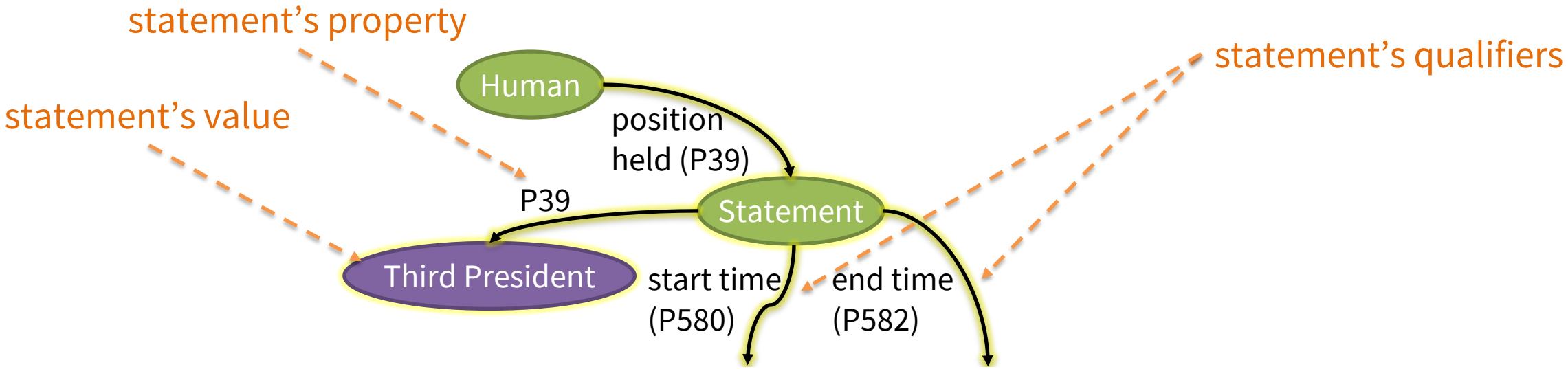


Name	Entered Office	Left Office	Party
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ

Third Presidents of National Council (Austria)

Source Modeling Problem

- Building semantic descriptions of tables
  - Describing data source using classes and properties in ontologies



Name	Entered Office	Left Office	Party
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ

Third Presidents of National Council (Austria)

Main Idea

- Information of entities in KGs can help source modeling

  $\Longrightarrow$ **need little training data**

President of the National Council (Austria)

From Wikipedia, the free encyclopedia

List of third presidents [edit]

Name	Entered Office	Left Office	Party
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ

Main Idea

- Information of entities in KGs can help source modeling

  ⟹ **need little training data**

President of the National Council (Austria)

From Wikipedia, the free encyclopedia

List of third presidents [edit]

Name	Entered Office	Left Office	Party
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ

Eva Glawischnig-Piesczek (Q93870)

Austrian politician edit

member of political party	Die Grünen	
position held	Third President of the National Council of Austria	
	start time	30 October 2006
	end time	28 October 2008

Main Idea

- Information of entities in KGs can help source modeling

  ⟹ **need little training data**

President of the National Council (Austria)

From Wikipedia, the free encyclopedia

List of third presidents [edit]

Name	Entered Office	Left Office	Party
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ

Eva Glawischnig-Piesczek (Q93870)

Austrian politician edit

member of political party	Die Grünen	
position held	Third President of the National Council of Austria	
	start time	30 October 2006
	end time	28 October 2008

Main Idea

- Information of entities in KGs can help source modeling

  ⟹ **need little training data**

President of the National Council (Austria)

From Wikipedia, the free encyclopedia

List of third presidents [edit]

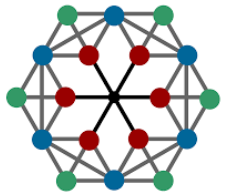
Name	Entered Office	Left Office	Party
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ

Eva Glawischnig-Piesczek (Q93870)

Austrian politician edit

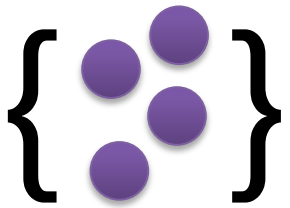
member of political party	Die Grünen
position held	Third President of the National Council of Austria
start time	30 October 2006
end time	28 October 2008

Approach

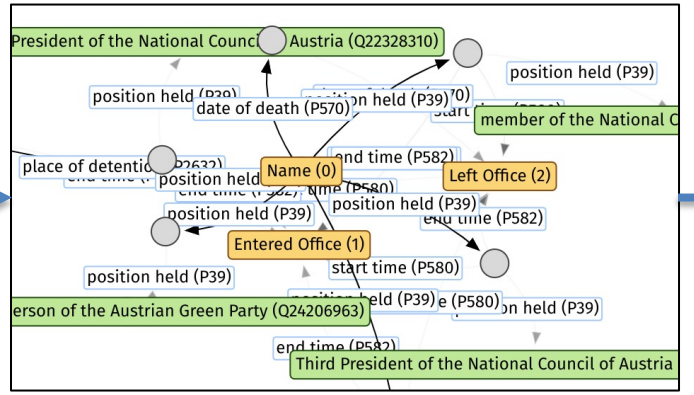


Name	Entered Office	Left Office	Party
Willi Brauneder	1996	1999	FPÖ
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ

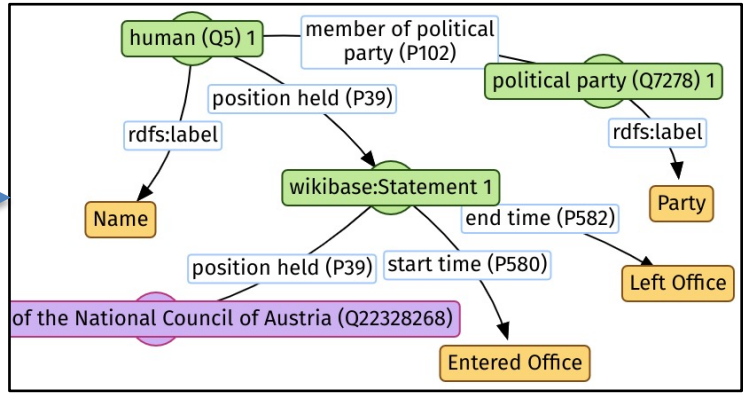
Linked table



Contextual values



Candidate Graph

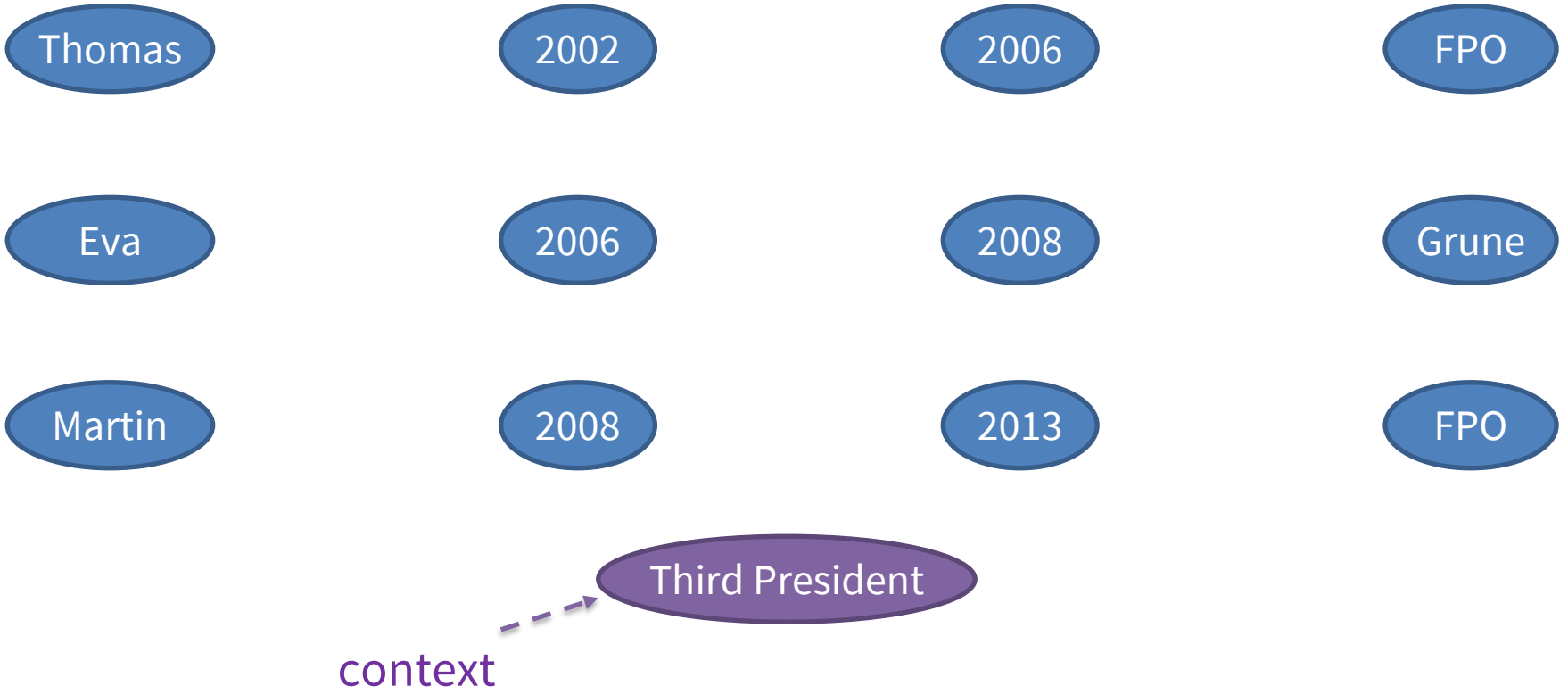


Semantic Description

Construct Candidate Graph: Discovering Links

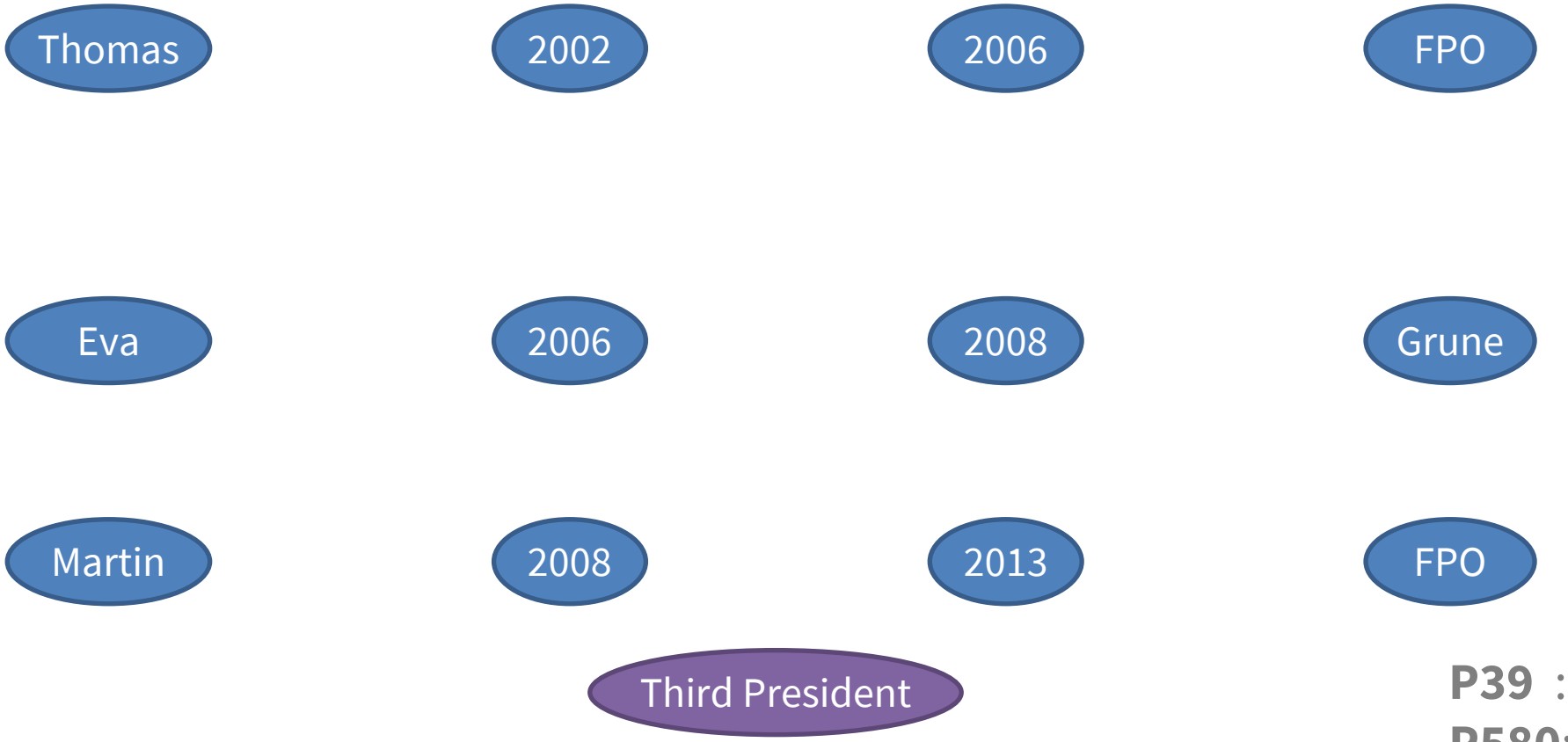
- Create a graph of cells and context

Name	Entered Office	Left Office	Party
Thomas Prinzhorn	2002	2006	FPÖ
Eva Glawischnig-Piesczek	2006	2008	Grüne
Martin Graf	2008	2013	FPÖ



Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata



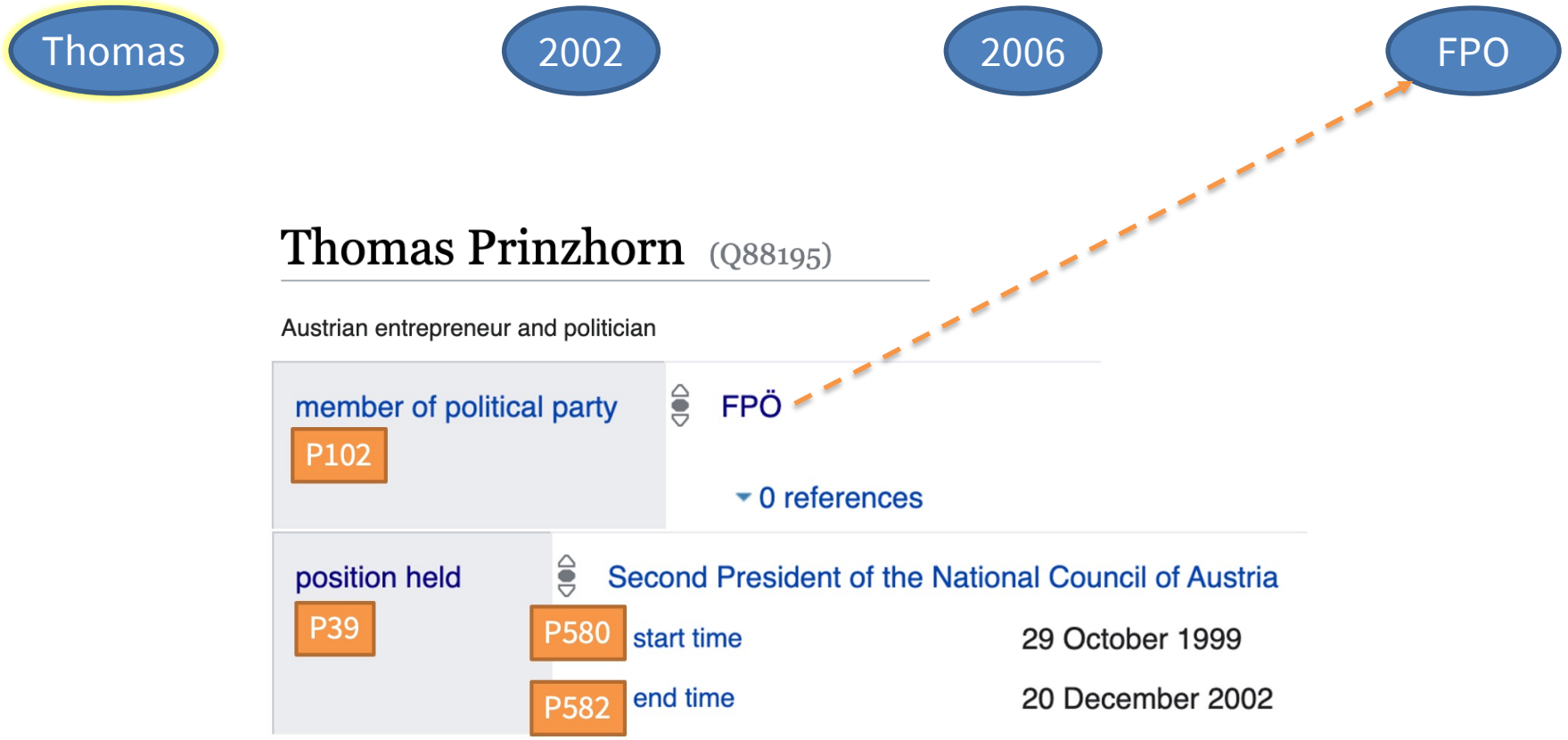
P39 : position held
P580: start time
P582: end time

Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata

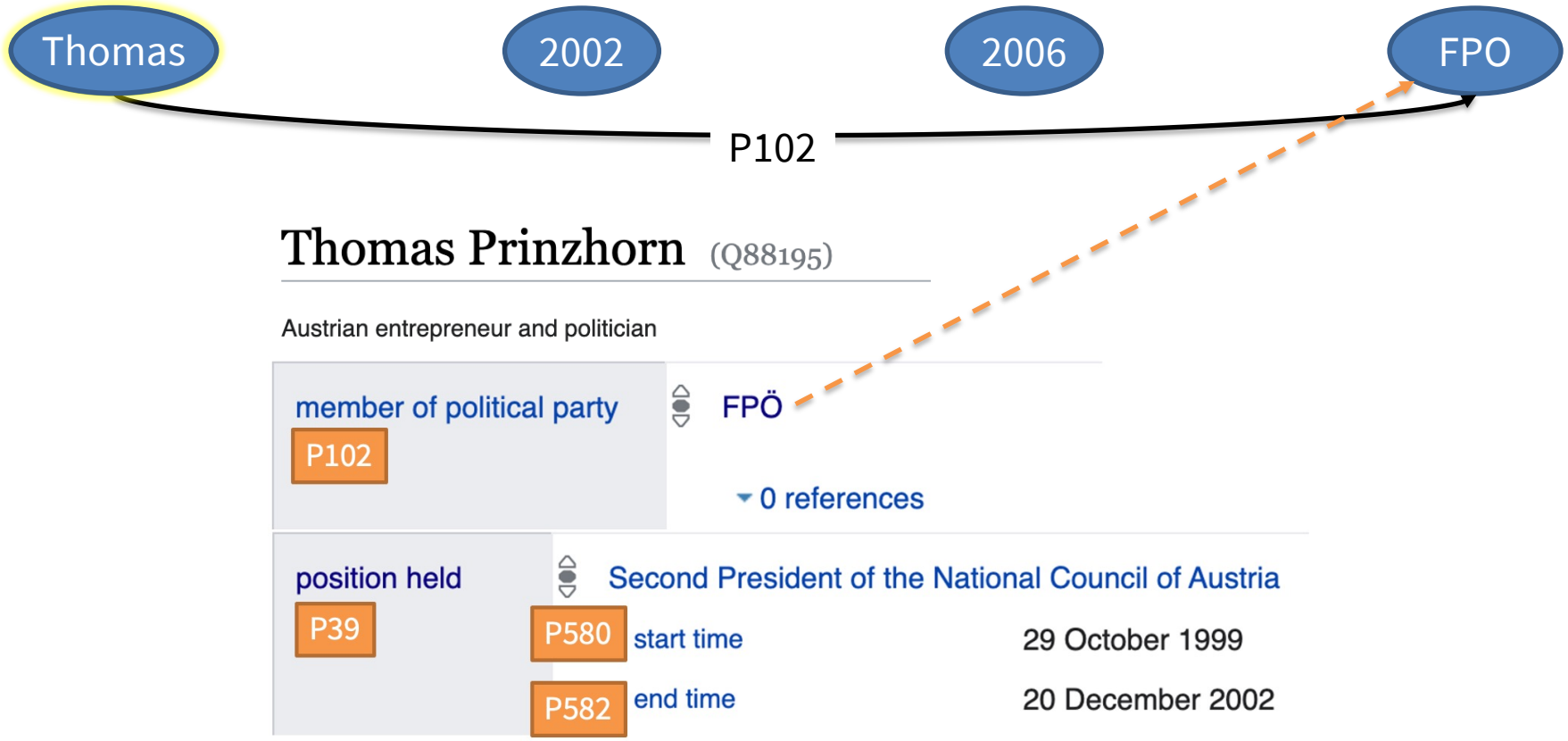
Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata



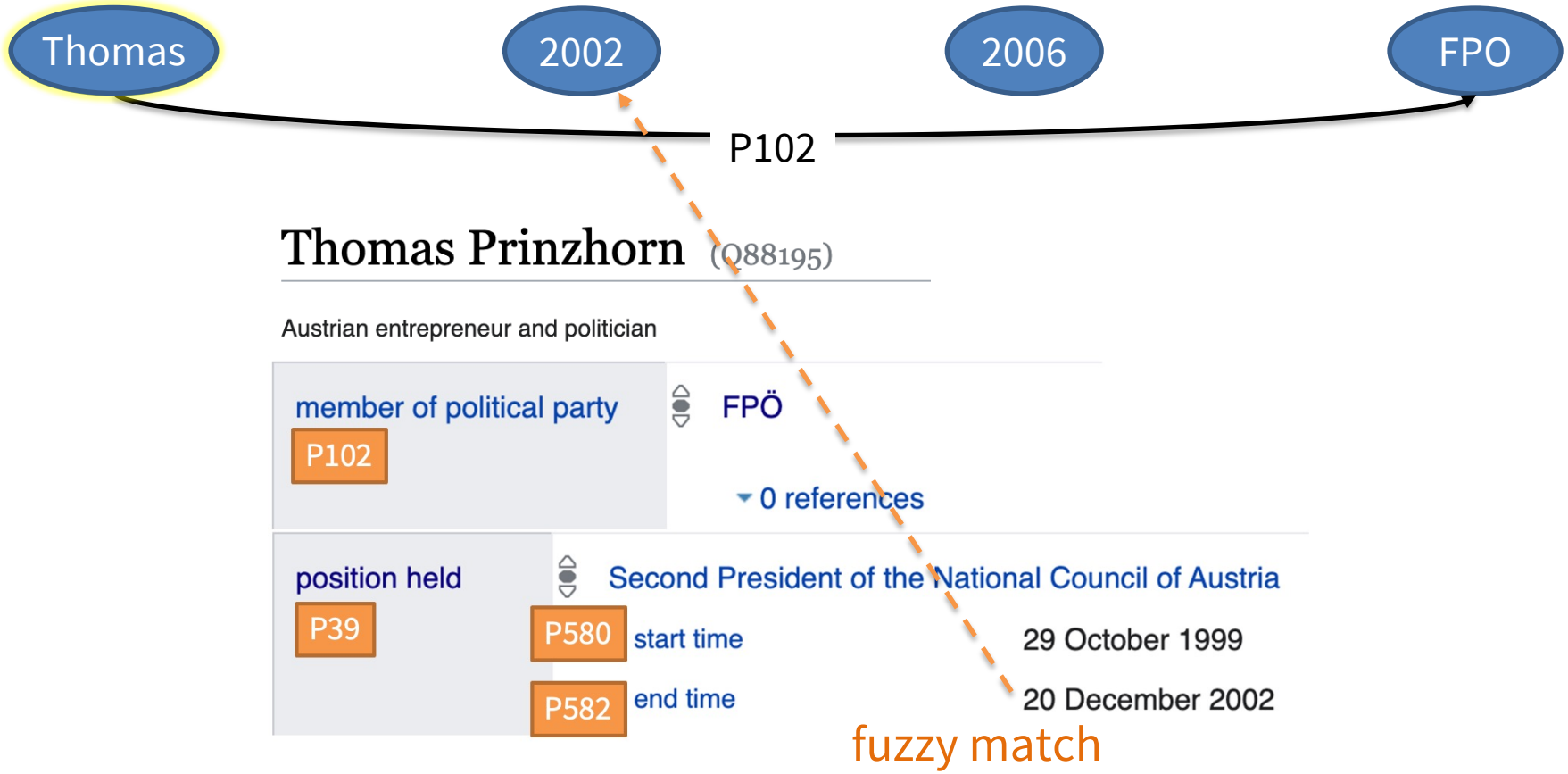
Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata



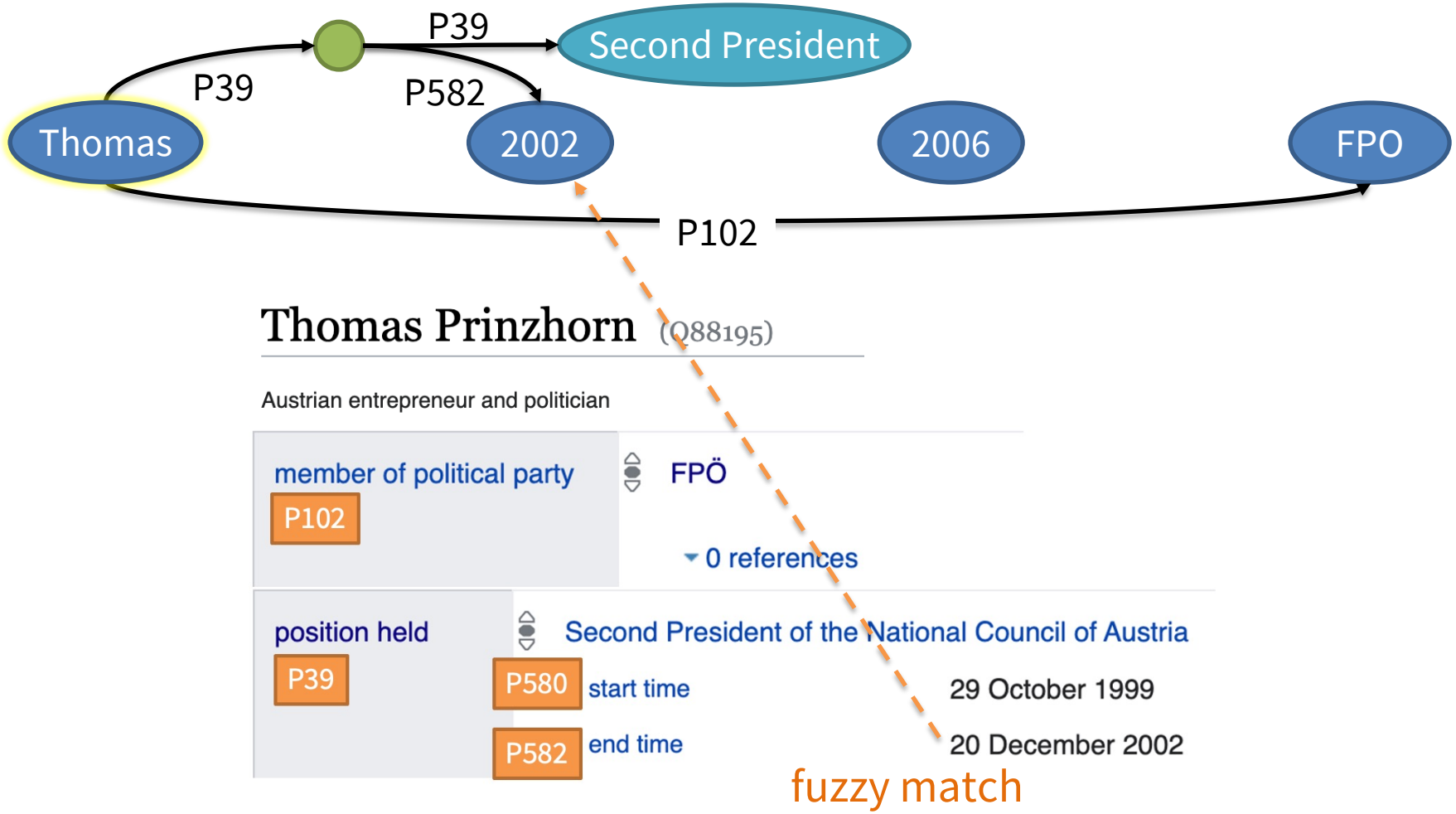
Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata



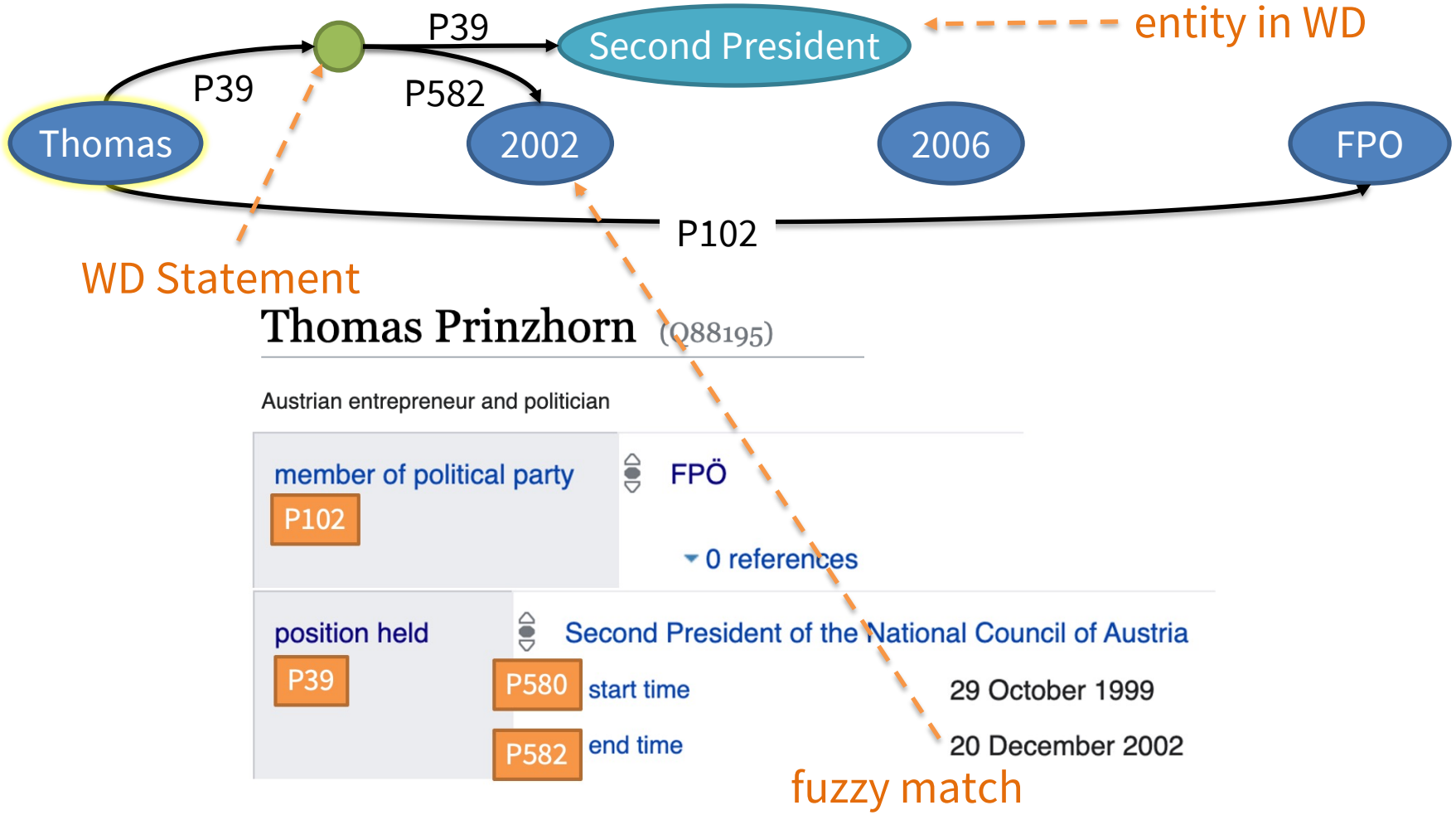
Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata



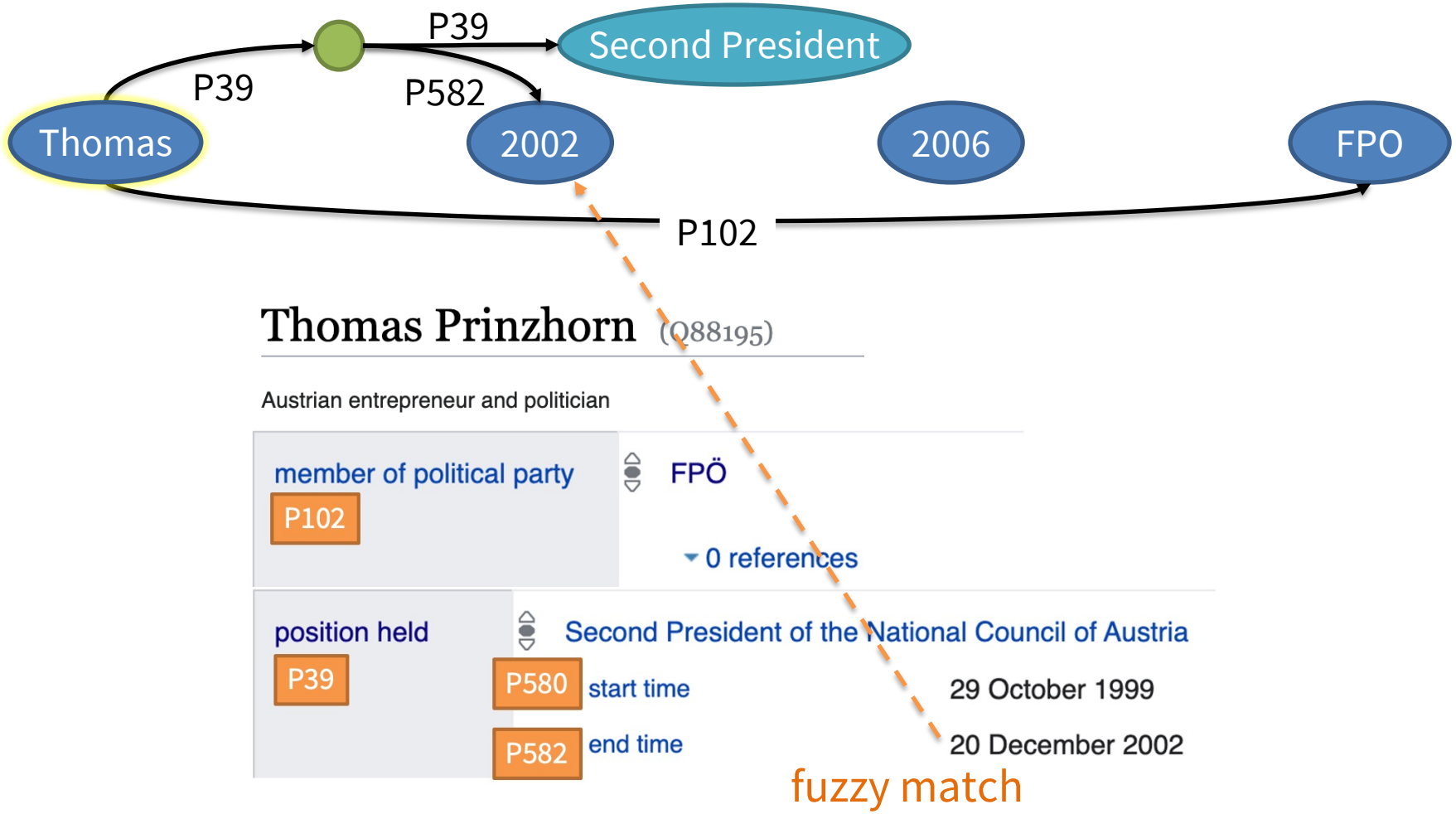
Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata



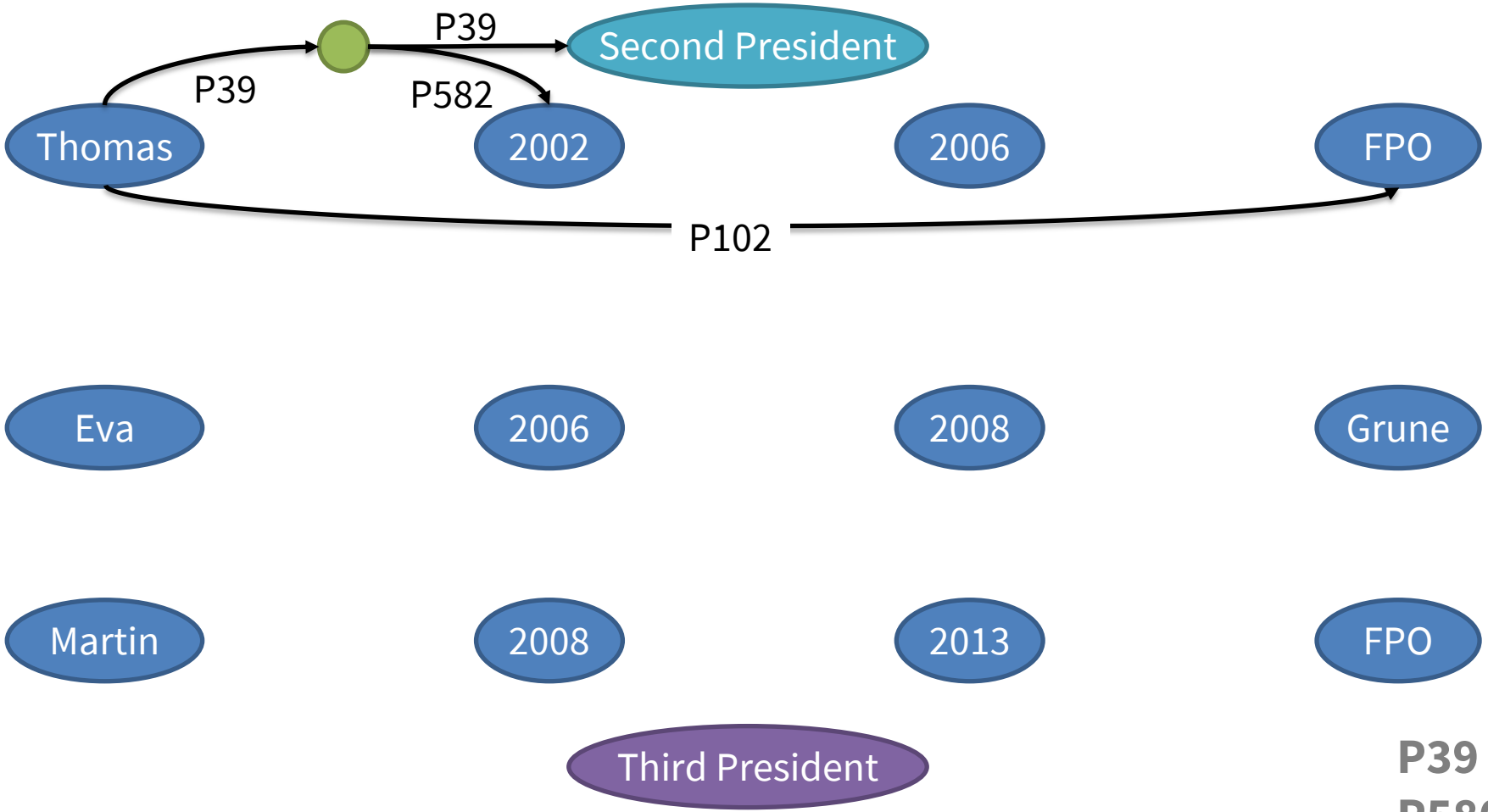
Construct Candidate Graph: Discovering Links

- Add links discovered from knowledge in Wikidata



Construct Candidate Graph: Discovering Links

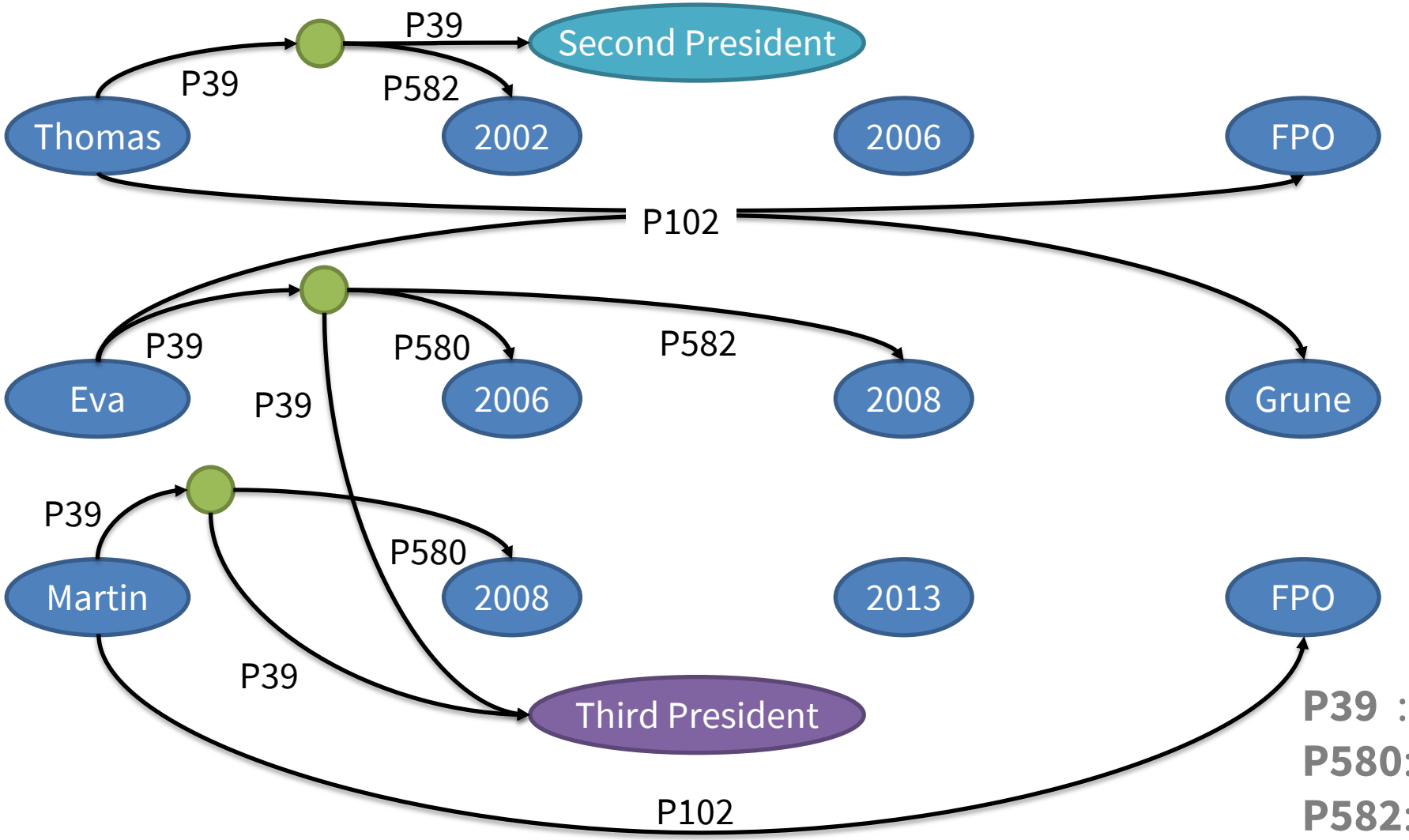
- Add links discovered from knowledge in Wikidata



P39 : position held
P580: start time
P582: end time

Construct Candidate Graph: Discovering Links

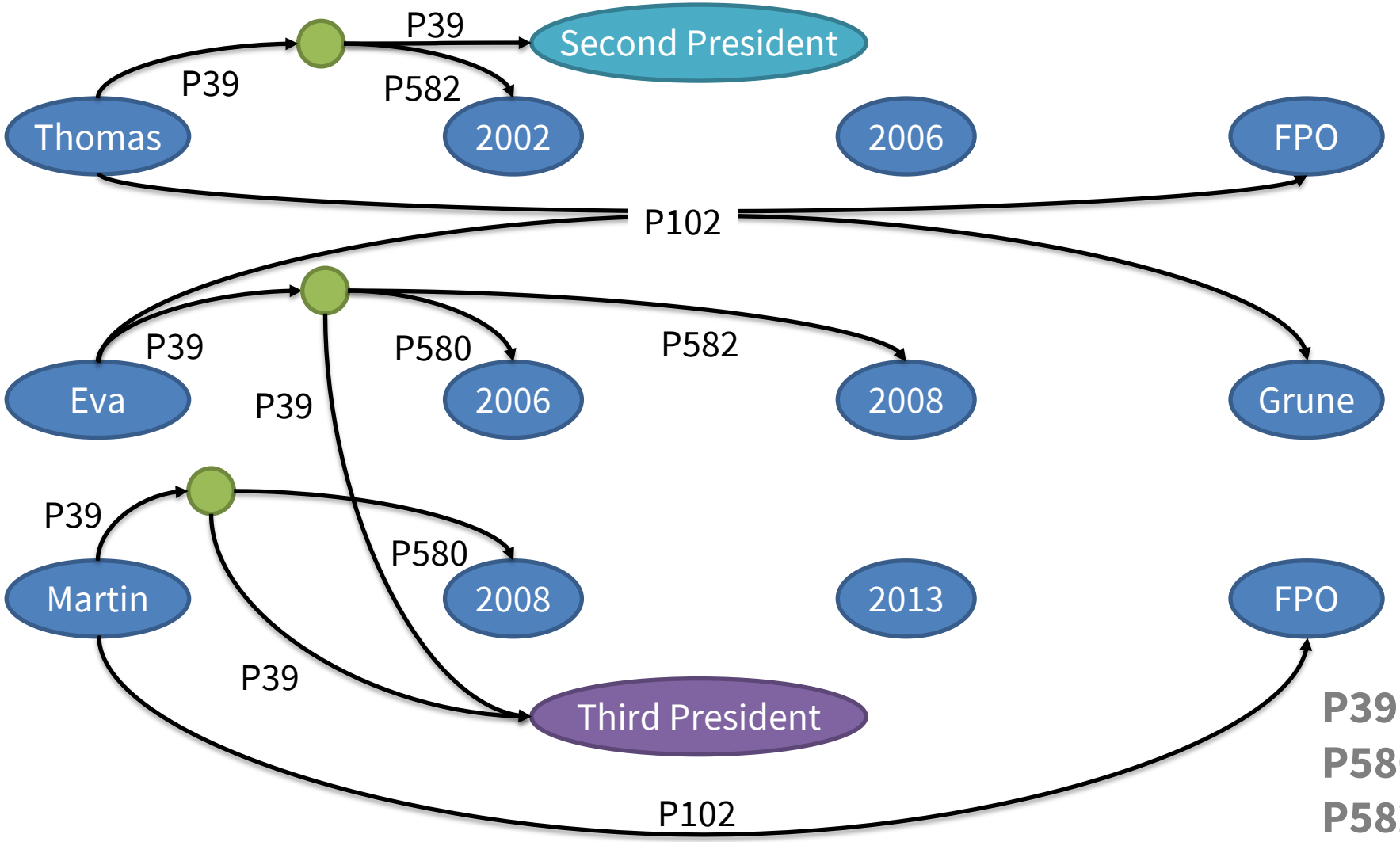
- Add links discovered from knowledge in Wikidata



P39 : position held
P580: start time
P582: end time

Construct Candidate Graph: Summarization

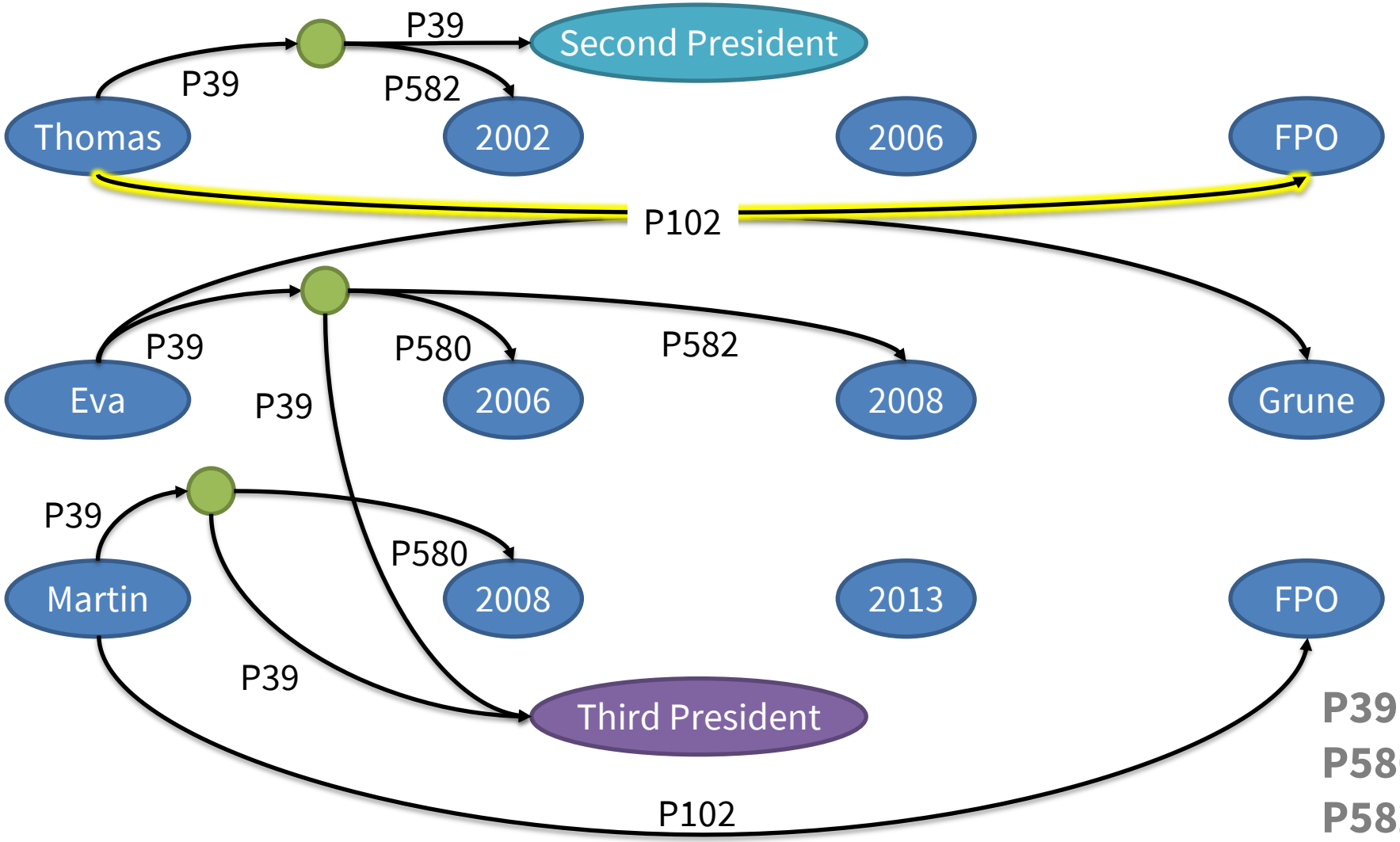
- Group links of cells from same source & target columns/context



P39 : position held
P580: start time
P582: end time

Construct Candidate Graph: Summarization

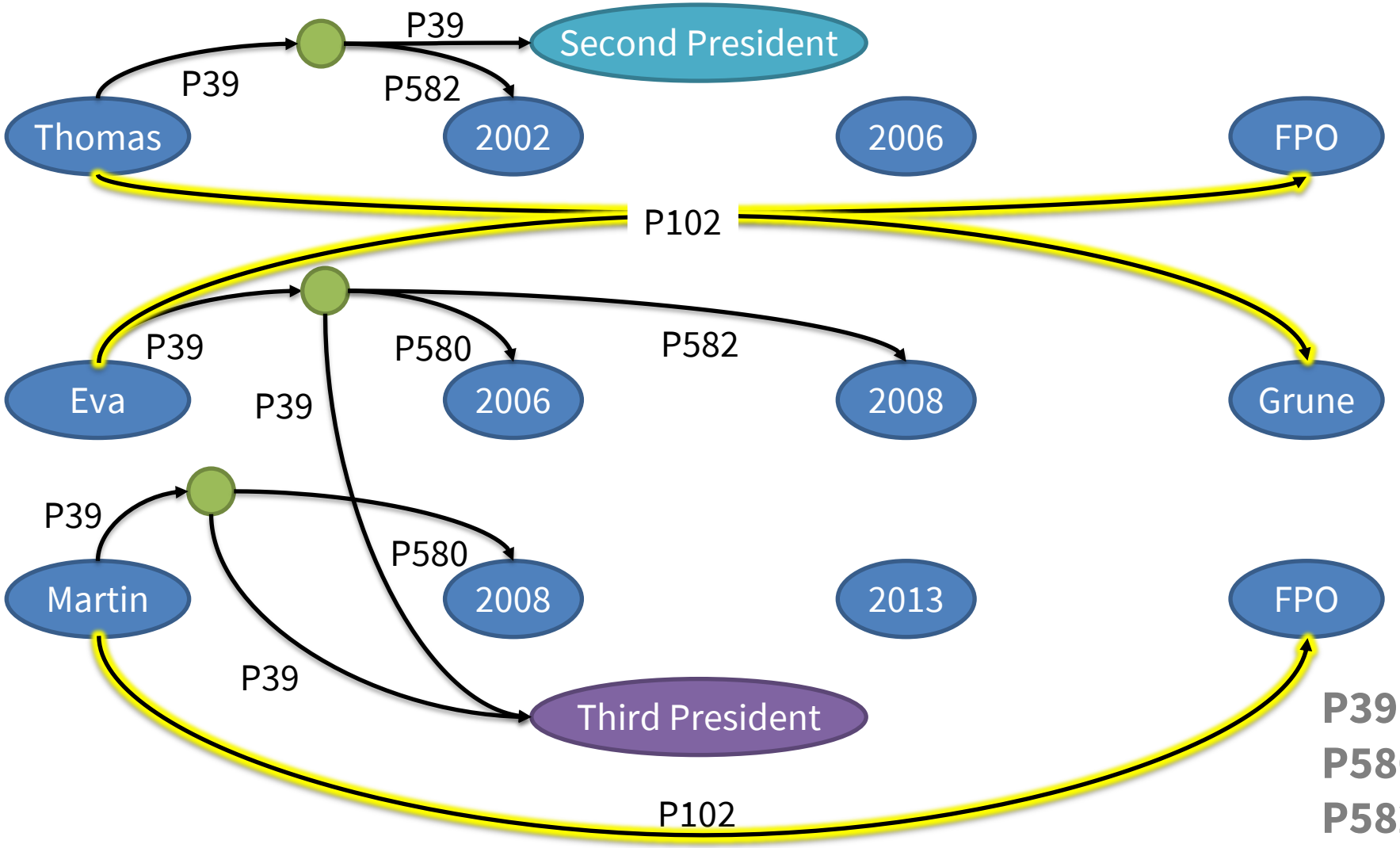
- Group links of cells from same source & target columns/context



P39 : position held
P580: start time
P582: end time

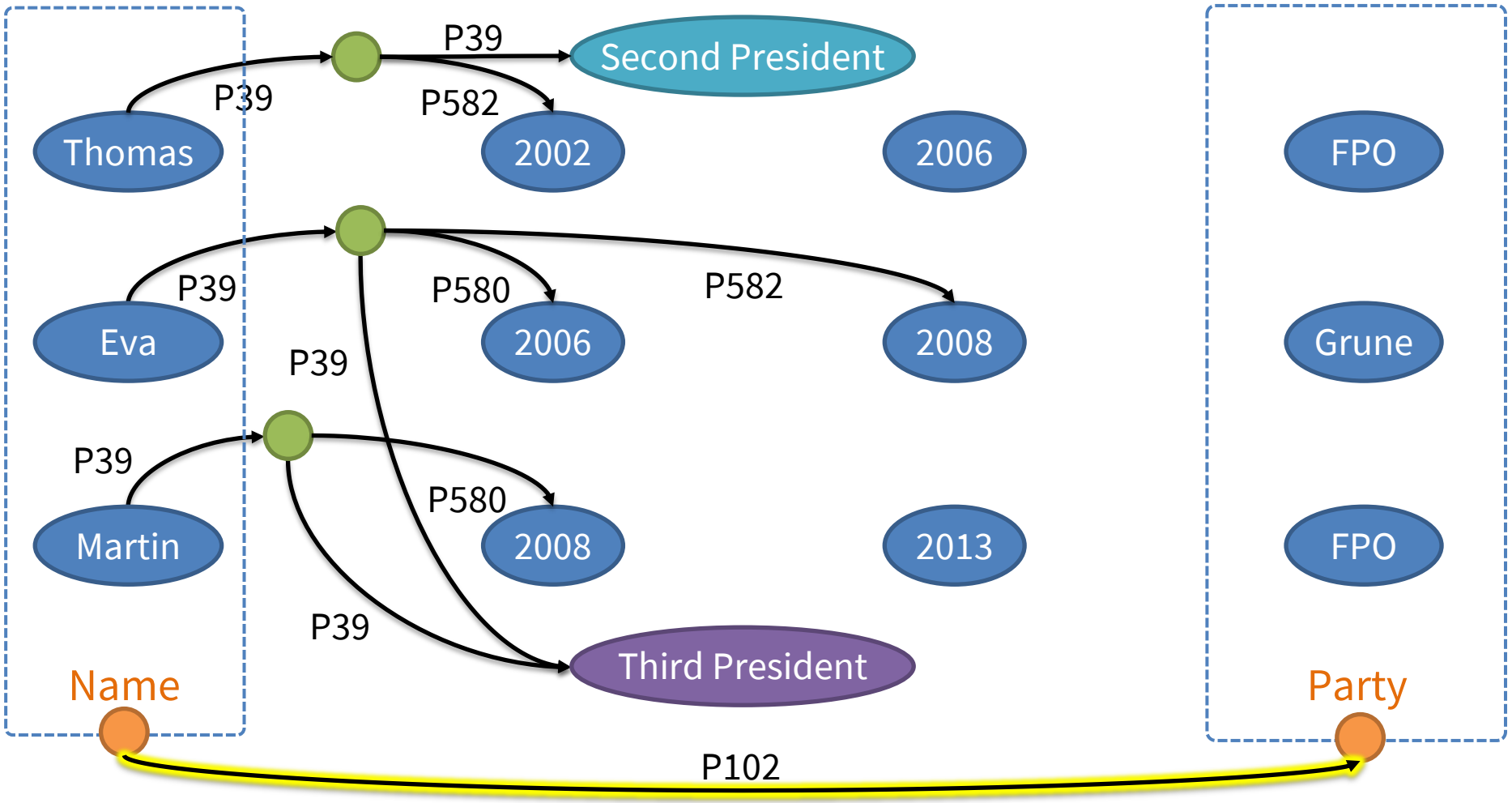
Construct Candidate Graph: Summarization

- Group links of cells from same source & target columns/context



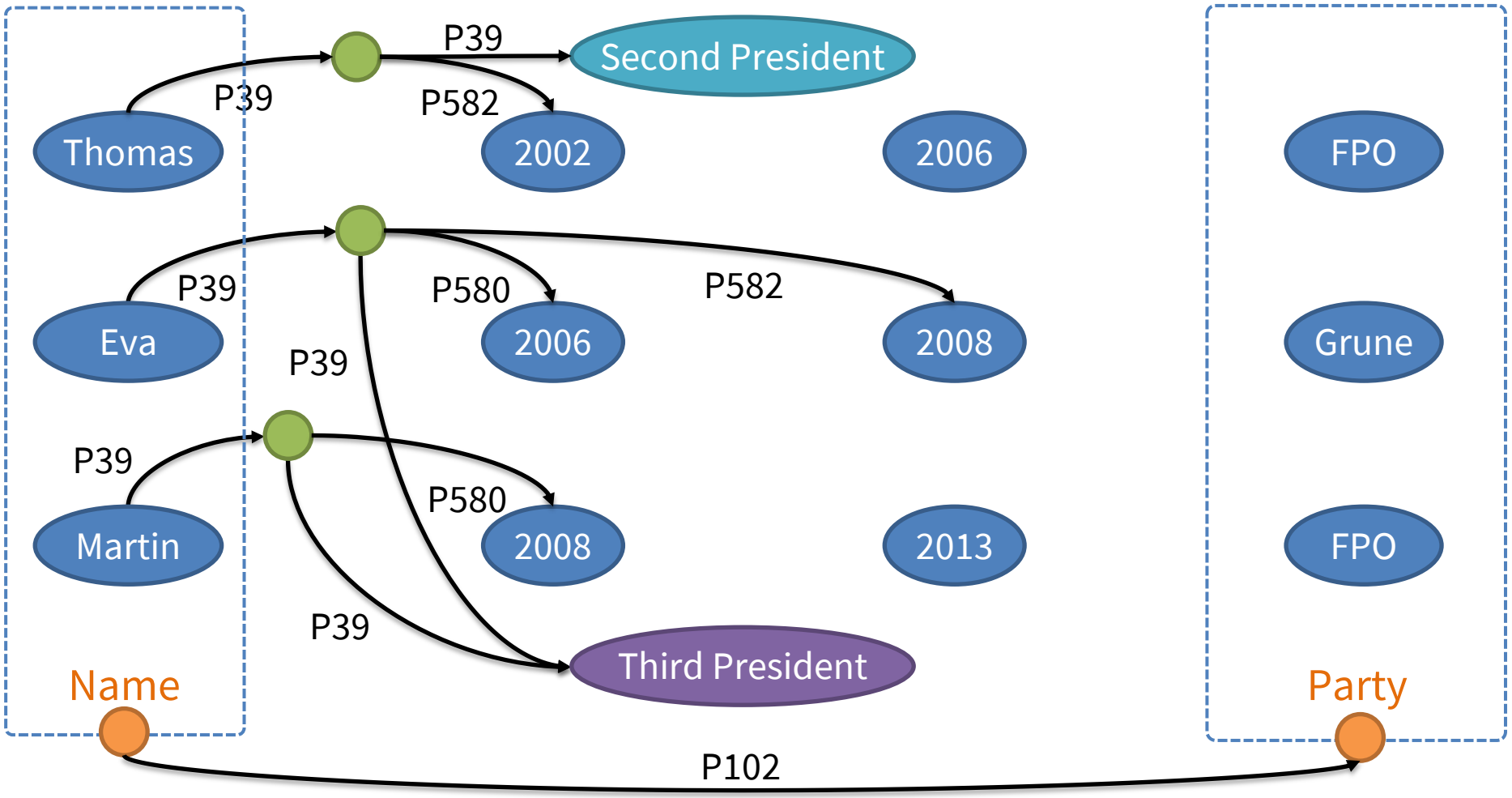
P39 : position held
P580: start time
P582: end time

Construct Candidate Graph: Summarization



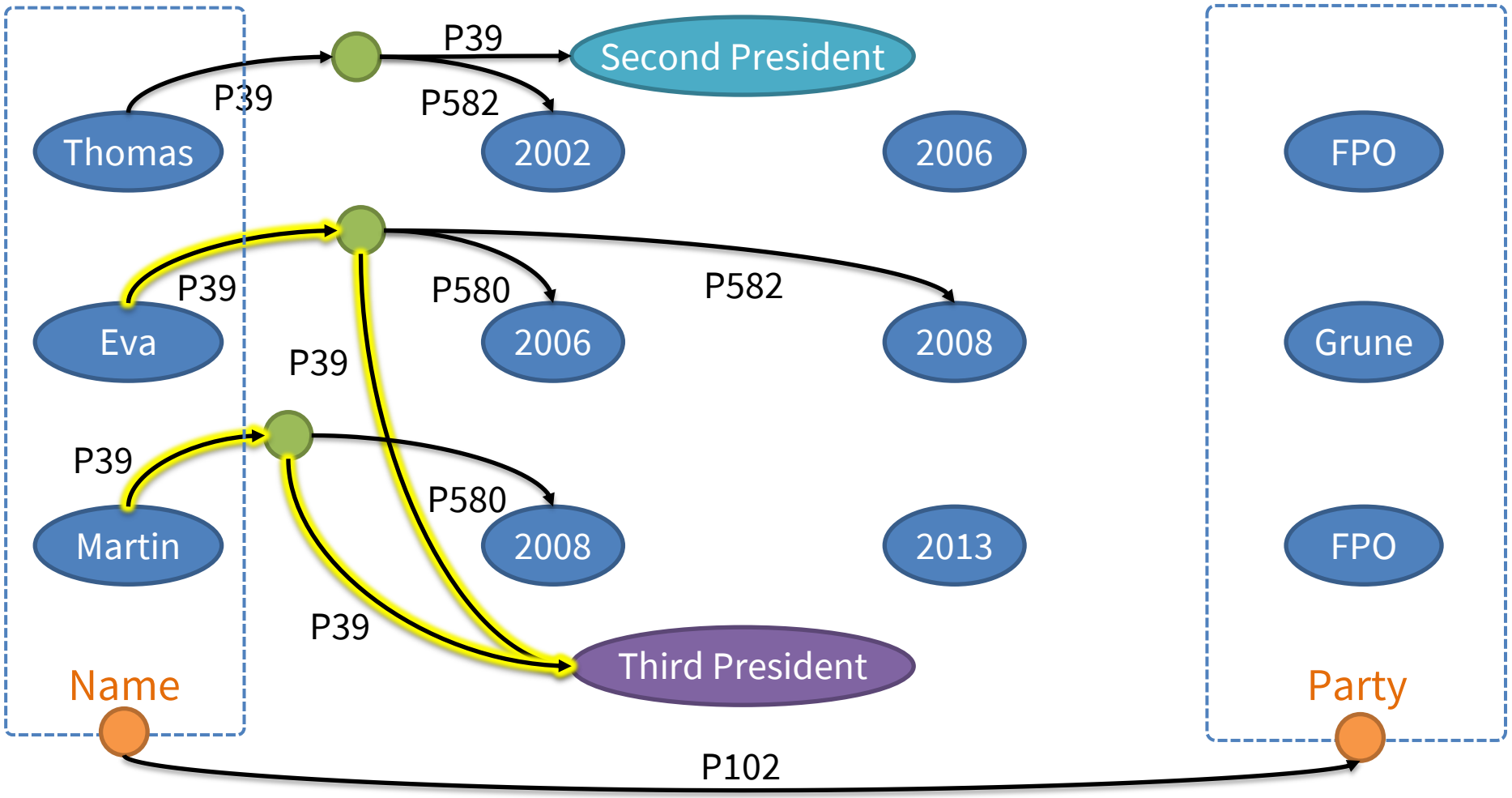
P39 : position held
P580: start time
P582: end time

Construct Candidate Graph: Summarization



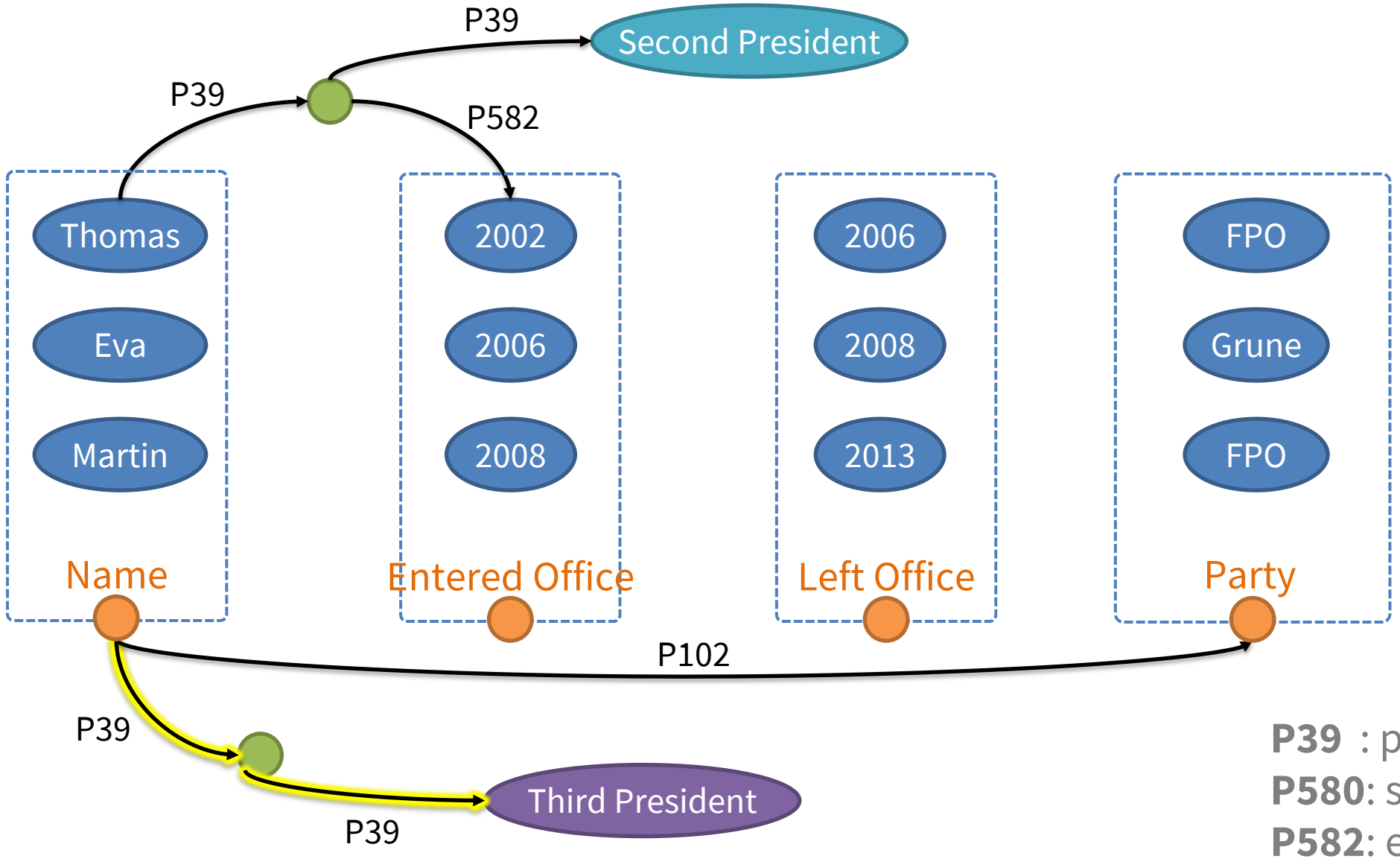
P39 : position held
P580: start time
P582: end time

Construct Candidate Graph: Summarization



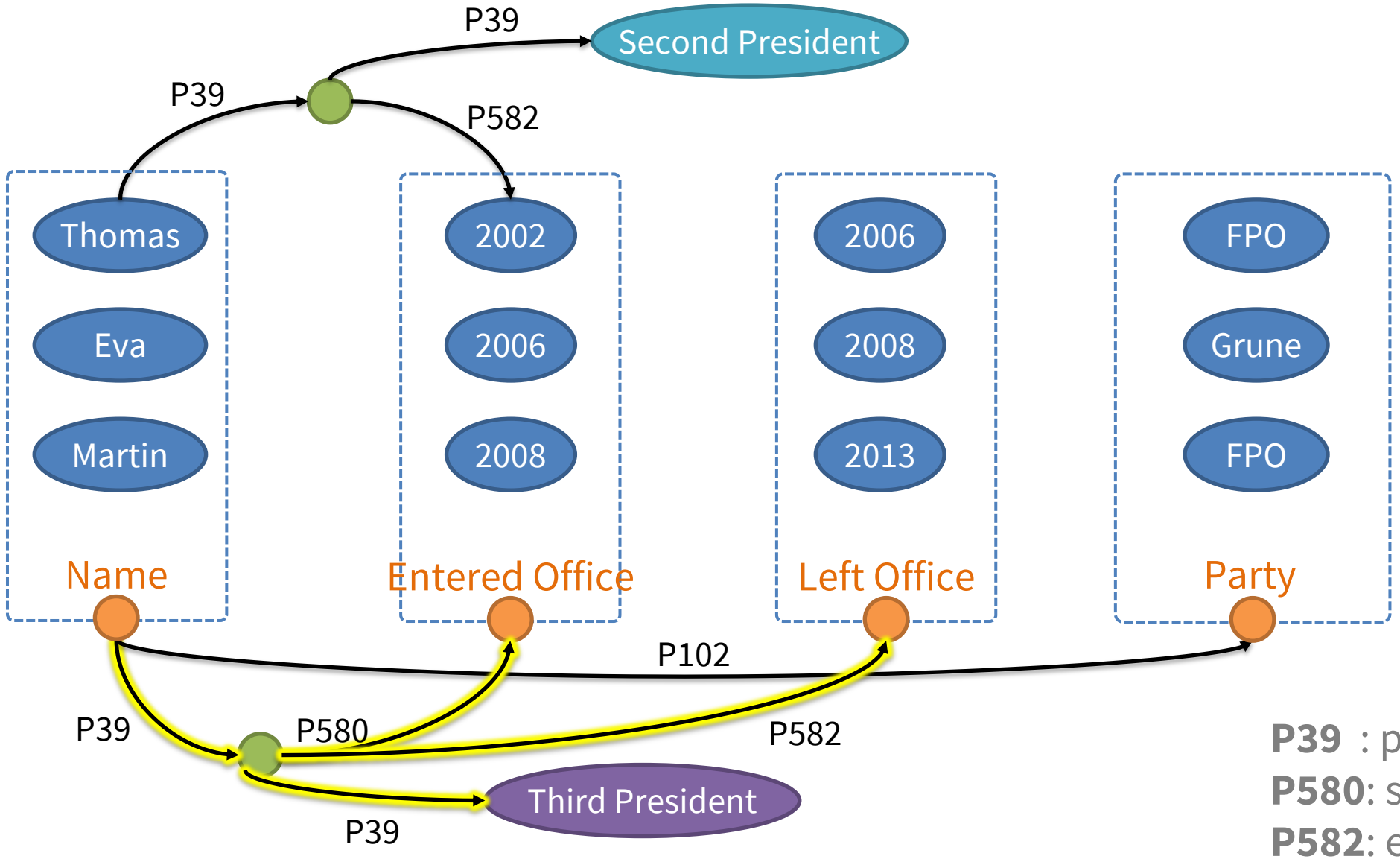
P39 : position held
P580: start time
P582: end time

Construct Candidate Graph: Summarization



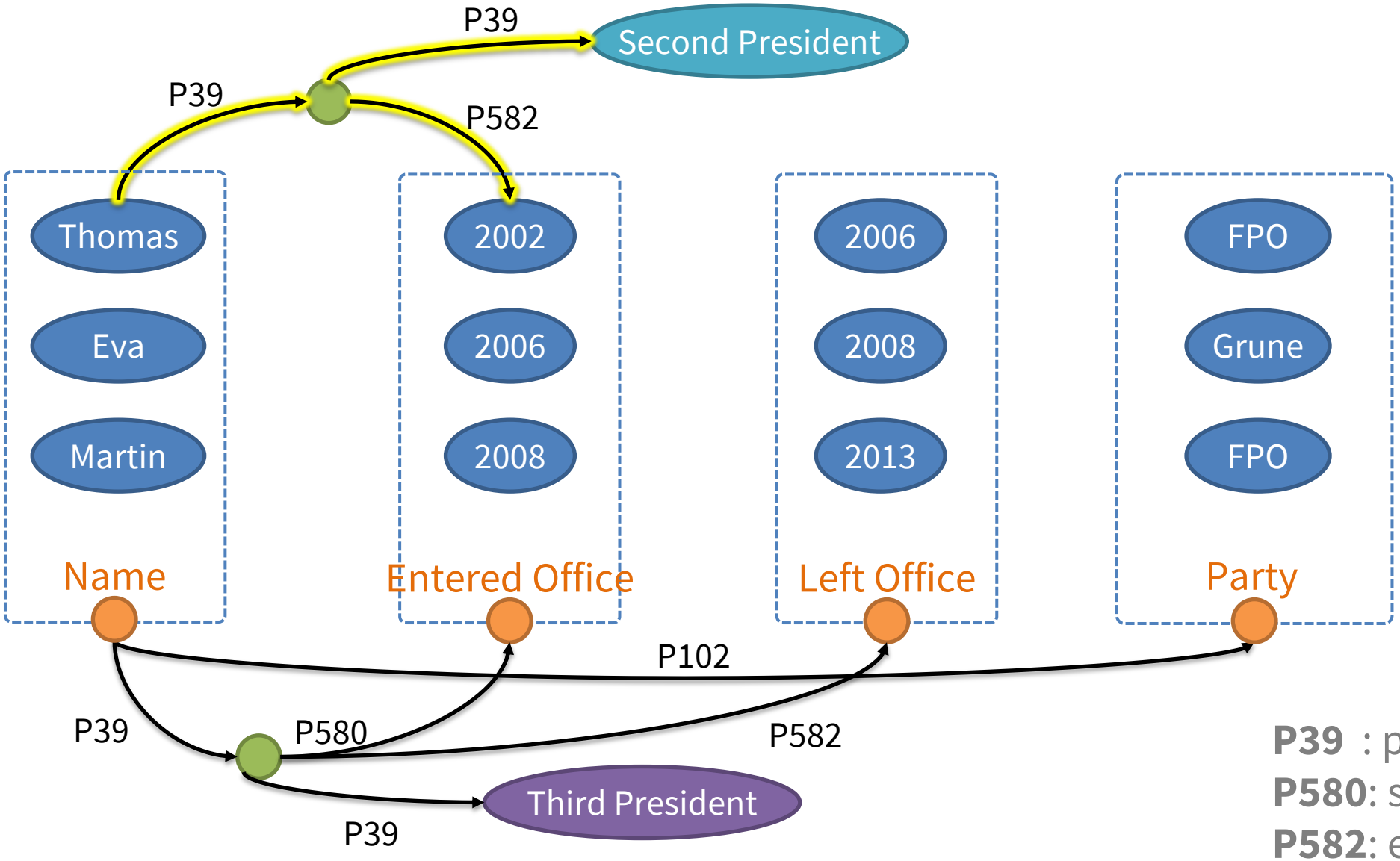
P39 : position held
P580: start time
P582: end time

Construct Candidate Graph: Summarization



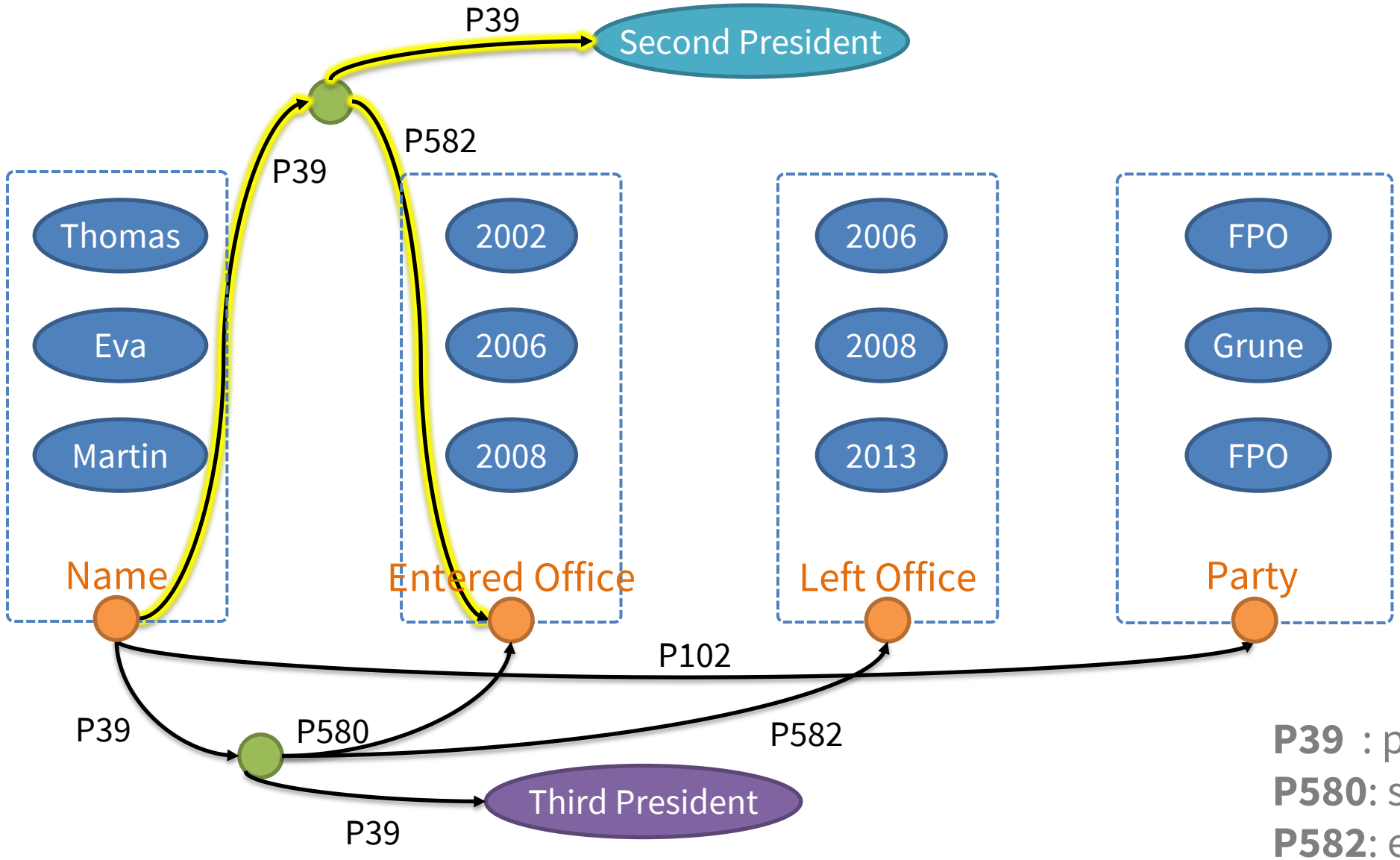
P39 : position held
P580: start time
P582: end time

Construct Candidate Graph: Summarization



P39 : position held
P580: start time
P582: end time

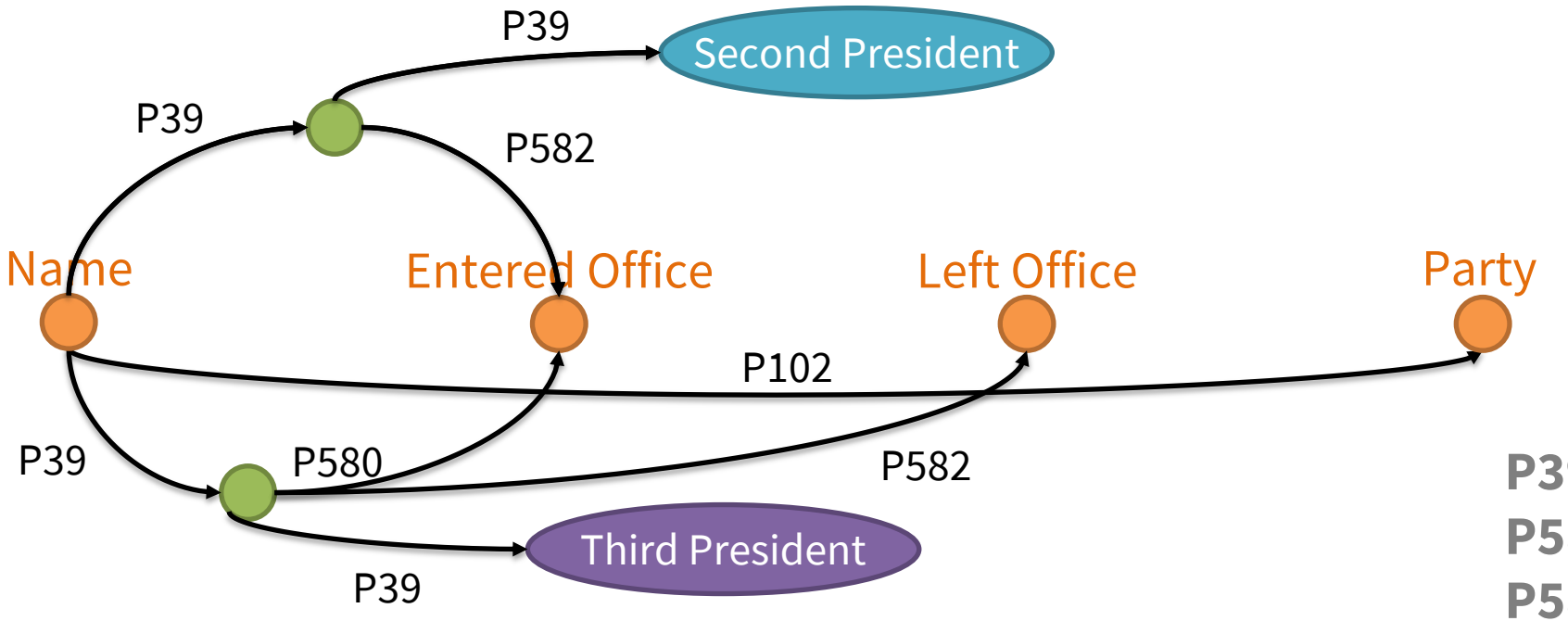
Construct Candidate Graph: Summarization



P39 : position held
P580: start time
P582: end time

Construct Candidate Graph: Summarization

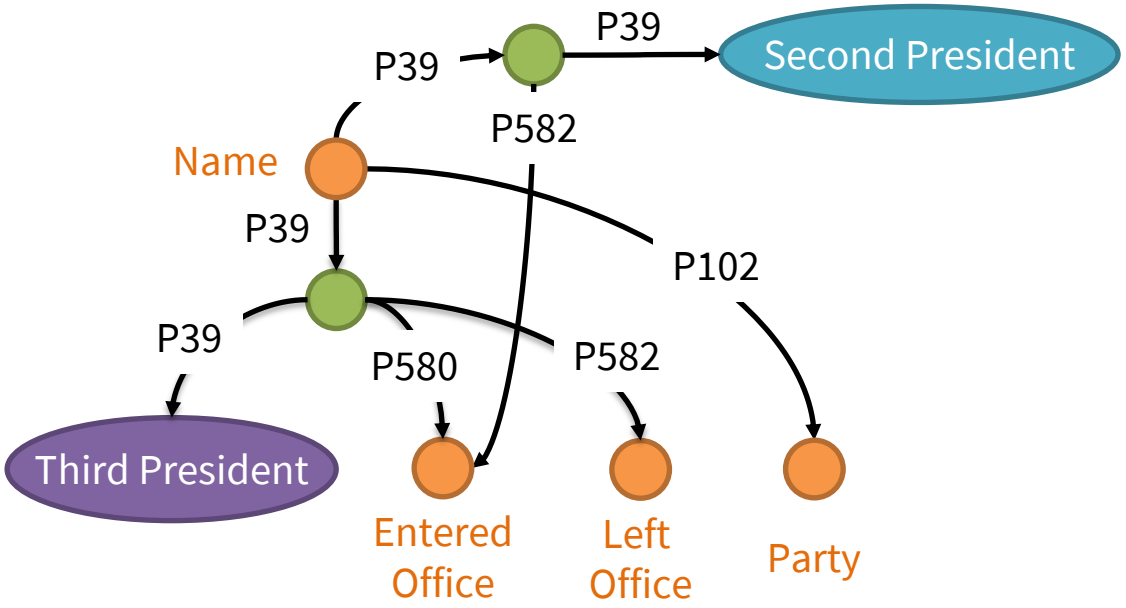
- Final candidate graph



P39 : position held
P580: start time
P582: end time

After Building Candidate Graph

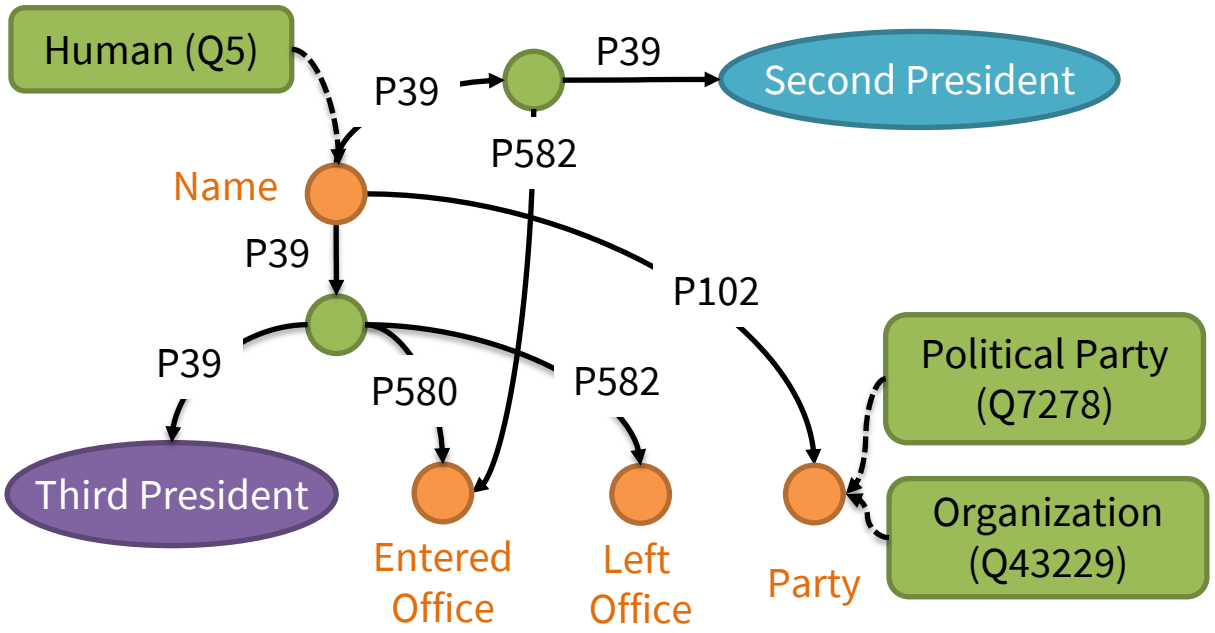
- Candidate (n-ary) relationships *from the candidate graph*



Candidate Graph

After Building Candidate Graph

- Candidate (n-ary) relationships *from the candidate graph*
- Candidate columns' types *from entities in table columns*



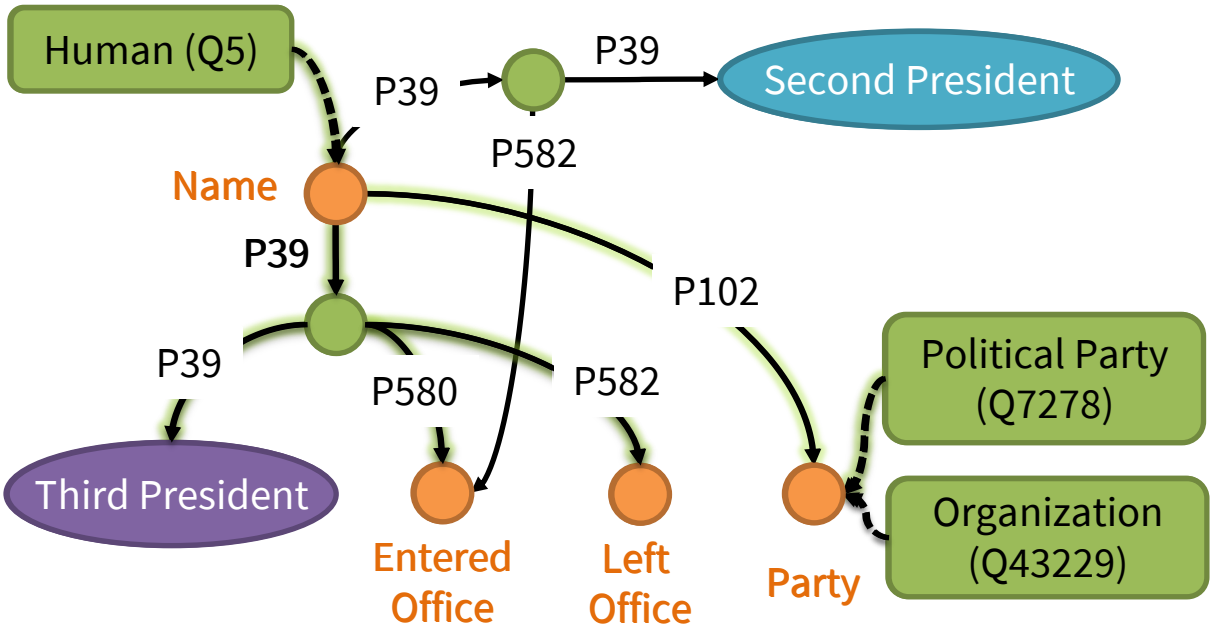
Semantic Description

Candidate Graph

After Building Candidate Graph

- Candidate (n-ary) relationships *from the candidate graph*
- Candidate columns' types *from entities in table columns*

⟹ Need to select the most appropriate relationships and types.



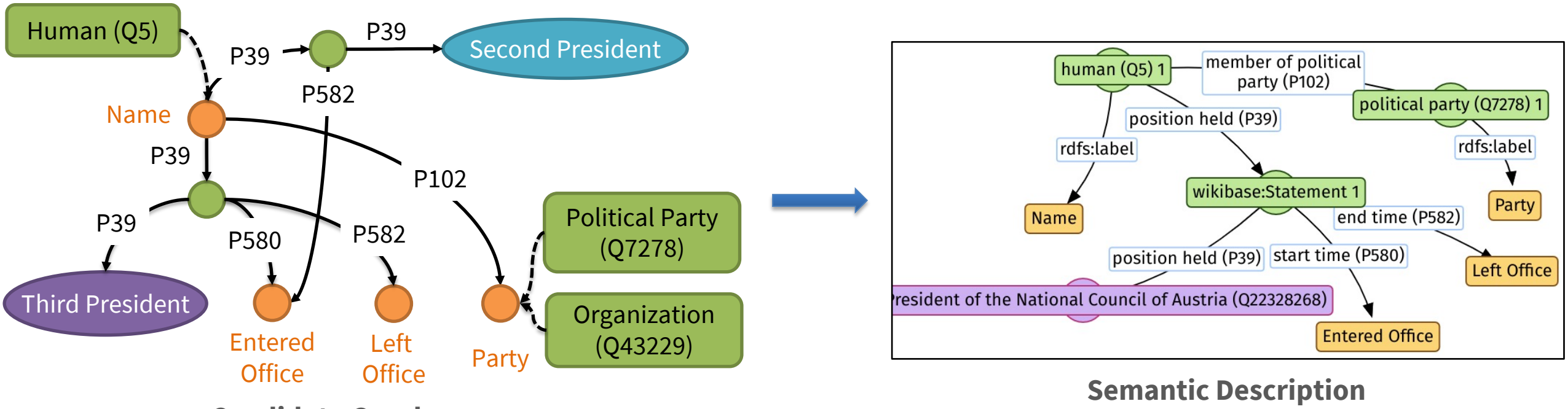
Candidate Graph

Semantic Description

After Building Candidate Graph

- Candidate (n-ary) relationships *from the candidate graph*
- Candidate columns' types *from entities in table columns*

⟹ Need to select the most appropriate relationships and types.



Candidate Graph

Semantic Description

Approach

Inputs

- A target knowledge graph: Wikidata
- A linked relational table T
- A set of contextual values C

1. Construct candidate graph
2. **Infer semantic description**

Outputs:

- A semantic description of (T, C)

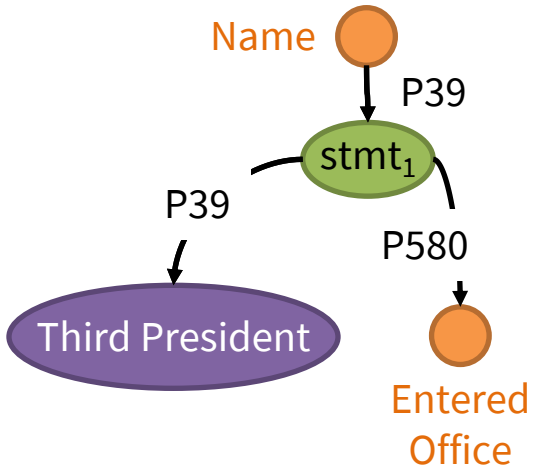
Collective Reasoning Problem

- **Probabilistic Soft Logic (PSL)**

  “A probabilistic graphical models framework using first-order logic”

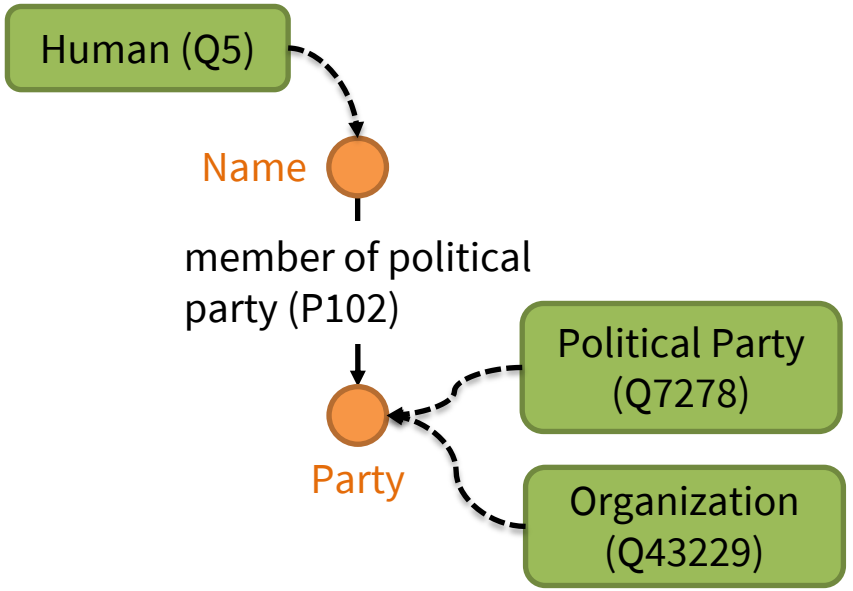
- Two main elements: **predicates** and **rules**
  - Predicates have “soft” value in [0, 1]
  - Rules converted to exponential function to approximate P(**x**)

PSL Predicates (examples)



- **CorrectRel(N_1, N_2, P)**: if a relationship is correct
  - CorrectRel(Name, $stmt_1$, P39)
  - CorrectRel($stmt_1$, Entered Office, P580)
  - CorrectRel($stmt_1$, Third President, P39)

- **CorrectType(N_1, T)**: if a column type assignment is correct
  - CorrectType(Party, Organization)
  - CorrectType(Party, Political Party)
  - CorrectType(Name, Human)



- … and more

P39: position held **P580**: start time **P582**: end time

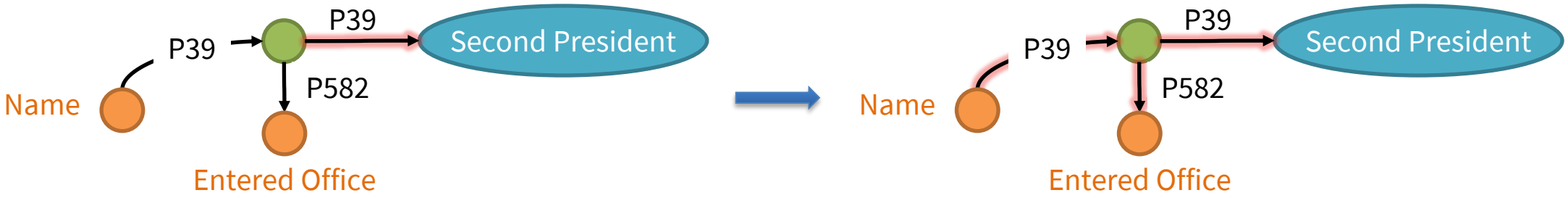
PSL Rules (examples)

1. By default, relationships/types are incorrect
   - 1a. $\neg$ CorrectRel(N_1, N_2, P)
   - 1b. $\neg$ CorrectType(N_1, T)

2. Relationships/types are correct/incorrect based on evidence
   - 2a. FreqMatch(N_1, N_2, P) $\rightarrow$ CorrectRel(N_1, N_2, P)
   - 2b. FreqDiff(N_1, N_2, P) $\rightarrow \neg$ CorrectRel(N_1, N_2, P)
   - 2c. FreqTypeMatch(N_1, T) $\rightarrow$ CorrectType(N_1, T)
   - 2d. ...and more

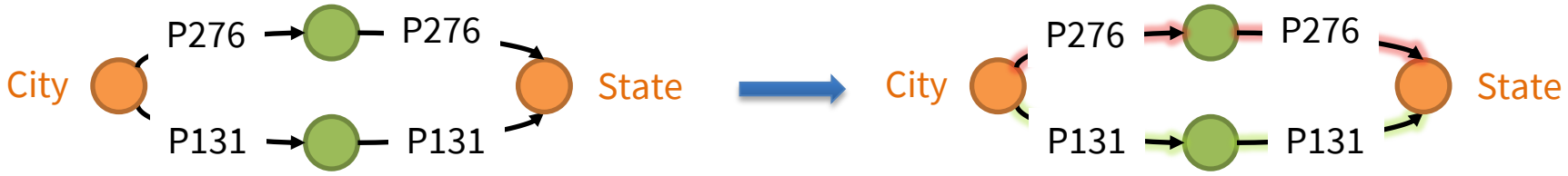
PSL Rules (examples)

3. If a statement value is incorrect, then the statement's qualifiers are also incorrect



4. We prefer fine-grain properties than high-level properties

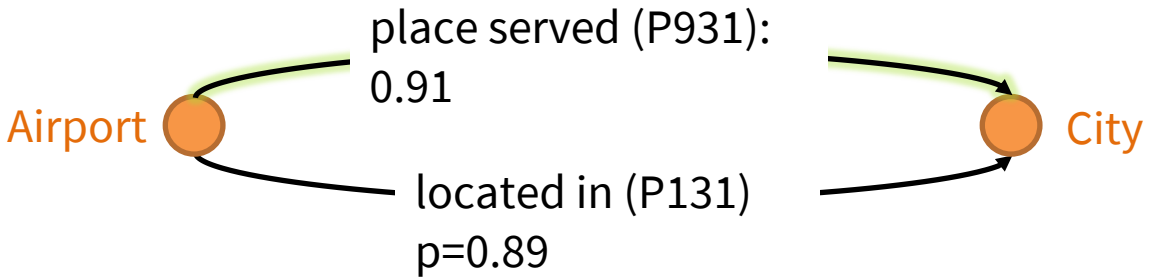
location (P276) is parent of **located in admin. area (P131)**



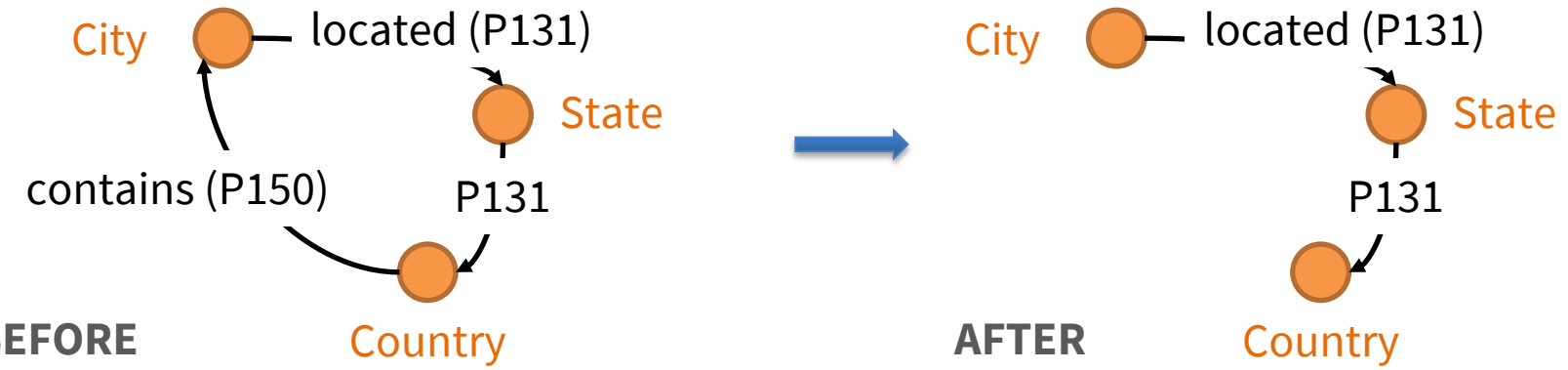
5. …and more

Post-Processing

- PSL outputs probability of each relationships and types.



- Use BANK algorithm to choose the most probable relationships
  - Avoid unnecessary loops
  - Prefer tree structure if possible



Evaluation of GRAMS

- Collective reasoning is beneficial
  - Avoid cascading errors from subject column detection phase
  - Handle complex schema: multiple entities' types and n-ary relationships

Wikipedia Tables

Synthetic Tables

Dataset	Method	CPA			CTA		
		Precision	Recall	F_1	Precision	Recall	F_1
250WT	MantisTable	0.535	0.442	0.484	0.928	0.331	0.488
	MantisTable*	0.559	0.569	0.564	**0.940**	0.394	0.556
	BBW	0.796	0.123	0.214	0.850	0.233	0.367
	BBW*	0.740	0.559	0.638	0.759	0.777	0.768
	GRAMS-ST	0.526	**0.681**	0.594	-	-	-
	GRAMS	**0.824**	0.650	**0.726**	0.819	**0.813**	**0.816**
SemTab2020	MantisTable	0.985	0.976	0.981	0.977	0.800	0.880
	BBW	**0.996**	**0.995**	**0.995**	0.980	0.980	0.980
	GRAMS-ST	0.990	0.989	0.990	-	-	-
	GRAMS	**0.996**	0.994	**0.995**	**0.982**	**0.981**	**0.982**

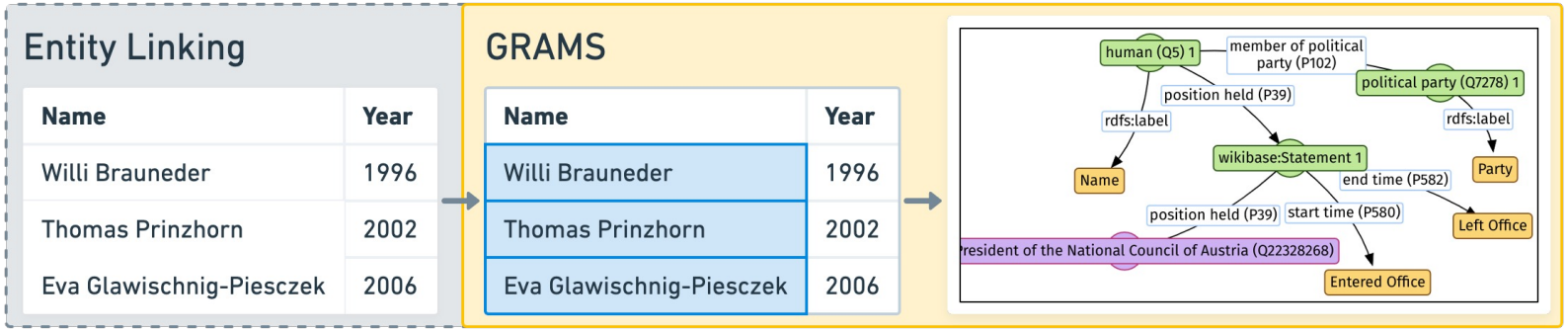
MantisTable* and BBW* are modified to retrieve correct subject column

Related Work

	Method		Data Hungry	Modeling Capabilities		
				Handle Literal Columns	Handle Qualifiers	Denormalized Tables
Custom Ontologies	Taheriyan et al. 2016		Y	Y	Y	Y
	Vu et al. 2019		Y	Y	Y	Y
KG Ontologies	Iterative Method	Ritze et al. 2015	-	Y	N	N
		Zhang et al. 2017	-	Y	N	N
		SemTab systems	-	Y	N	N
	Graphical Models	Limaye et al. 2010	-	N	N	Y
		Mulward et al. 2013	-	N	N	Y
		GRAMS	-	Y	Y	Y

Discussion and Future work

- **Contribution:** A novel graph-based approach, GRAMS, for building semantic descriptions of Wikipedia Tables.
  - The candidate graph makes it easy to represent and discover n-ary relationships.
  - Using PSL to collectively infer correct relationships and types.
- Future work:
  - Handle unlinked tables



  - Generate large labeled dataset from Wikipedia tables to train semantic modeling systems