

Finding Name Co-occurrences in WDY

Elif Yamangil & Rani Nelken

**What are name
co-occurrences?**

What are name co-occurrences?

X	
Y	
Z	
	...
	...

What are name co-occurrences?

X	Y
Y	
Z	
	...
	...

What are name co-occurrences?

X	Y	Z
Y		
Z		
		...
		...

What are name co-occurrences?

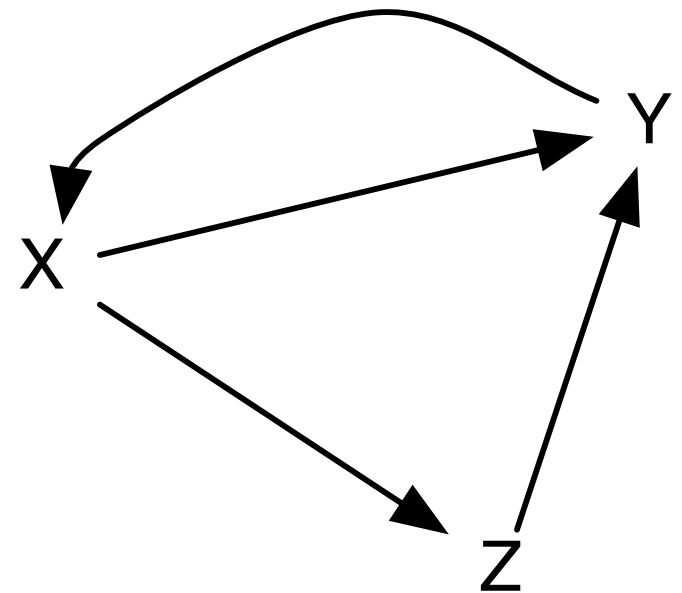
X	Y	Z
Y	X	
Z		
		...
		...

What are name co-occurrences?

X	Y	Z
Y	X	
Z	Y	
	...	
	...	

What are name co-occurrences?

X	Y	Z
Y	X	
Z		Y
	...	
	...	



How do we find them?

- Simple search!
- Collect all names X,Y, Z, ...
- Iterate through all biographies
- Search for the names
- Report the longest earliest match

Example

Example

- 王翦 字一飛，成都人。篤學尚氣。吳曦謀反來請，翦陽病風瘡，潛往安丙，謀誅曦。事定，匿巴中，為農終身。

Example

- 王翦 字一飛，成都人。篤學尚氣。吳曦謀反來請，翦陽病風瘡，潛往安丙，謀誅曦。事定，匿巴中，為農終身。

Example

- 王翦 字一飛，成都人。篤學尚氣。吳曦謀反來請，翦陽病風瘡，潛往安丙，謀誅曦。事定，匿巴中，為農終身。

Example

- 王翥 字一飛，成都人。篤學尚氣。吳曦謀反來請，翥陽病風瘖，潛往安丙，謀誅曦。事定，匿巴中，為農終身。
- 丁偃 吳縣人。嘉祐四年與朱長文同登進士。其初試邇英延講藝有詩云：「白虎前芳掩，金華舊事經，天心非不寢，垂意在蒼生。」御前下第，後二十年方中選。

Example

- 王翥 字一飛，成都人。篤學尚氣。吳曦謀反來請，翥陽病風瘖，潛往安丙，謀誅曦。事定，匿巴中，為農終身。
- 丁偃 吳縣人。嘉祐四年與朱長文同登進士。其初試邇英延講藝有詩云：「白虎前芳掩，金華舊事經，天心非不寢，垂意在蒼生。」御前下第，後二十年方中選。

Example

- 王翥 字一飛，成都人。篤學尚氣。吳曦謀反來請，翥陽病風瘖，潛往安丙，謀誅曦。事定，匿巴中，為農終身。
- 丁偃 吳縣人。嘉祐四年與朱長文同登進士。其初試邇英延講藝有詩云：「白虎前芳掩，金華舊事經，天心非不寢，垂意在蒼生。」御前下第，後二十年方中選。

Example

- 王翥 字一飛，成都人。篤學尚氣。吳曦謀反來請，翥陽病風瘖，潛往安丙，謀誅曦。事定，匿巴中，為農終身。
- 丁偃 吳縣人。嘉祐四年與朱長文同登進士。其初試邇英延講藝有詩云：「白虎前芳掩，金華舊事經，天心非不寢，垂意在蒼生。」御前下第，後二十年方中選。

Problem:

Name Ambiguity

X	Y	Z
Y	X	
Z		Y
		...
		...

Problem:

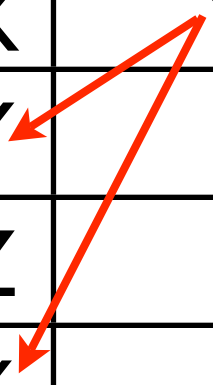
Name Ambiguity

X	Y	Z
Y	X	
Z		Y
Y		...
		...

Problem:

Name Ambiguity

X	Y	Z
Y	X	
Z	Y	
Y	...	
	...	



The diagram illustrates name ambiguity using a table. The table has five rows and three columns. The first column contains the names X, Y, Z, Y, and an empty cell. The second column contains Y, X, Y, ..., and ... The third column contains Z, an empty cell, an empty cell, an empty cell, and an empty cell. Two red arrows originate from the 'Y' in the first column of the first row and point to the 'Y' in the second column of the second row and the 'Y' in the second column of the third row, highlighting the ambiguity of the name 'Y'.

Problem:

Name Ambiguity

X	Y	Z
Y	X	
Z		Y
Y	...	
	...	

The diagram illustrates name ambiguity using a table with five rows and three columns. Red arrows point from the 'Y' entries in the first column to the 'Y' entries in the second and third columns, showing that the same name 'Y' is used in different contexts, leading to ambiguity.

Problem: Name Ambiguity

X	Y	Z
Y	X	
Z		Y
Y	...	
	...	

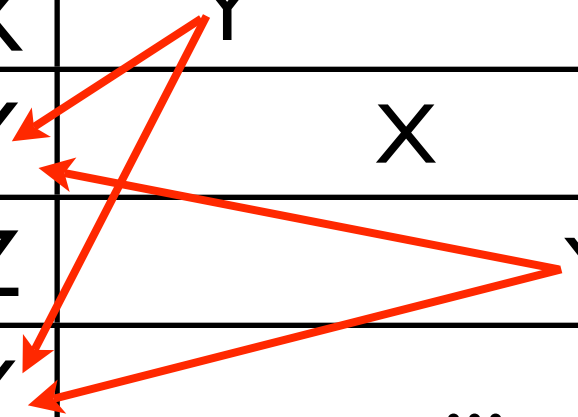
?

Example

丁騭	丁騭 字公點，武進人。嘉祐二年進士。 李定 用事，辟為屬，以疾辭。蘇軾、曾肇、孔文仲交薦之，除太常博士，改右正言。元祐間士風險兢，有五鬼十物之號，騭疏請窮治。後出知處州。騭以經學倡後進，長於易春秋，為文自成一家。有文集二十卷。
李定	...
李定	...
李定	...
李定	...

Disambiguation

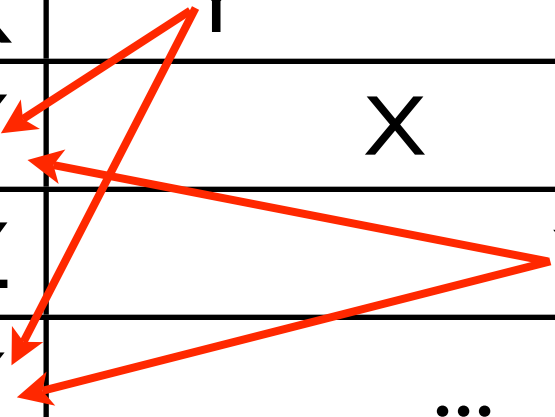
X	Y	Z
Y	X	
Z		Y
Y	...	
	...	



The diagram illustrates the disambiguation of a table. Red arrows point from the 'Y' entries in the first column to the 'Y' entries in the other columns, indicating that the 'Y' in the first column is distinct from the 'Y' in the other columns.

Disambiguation

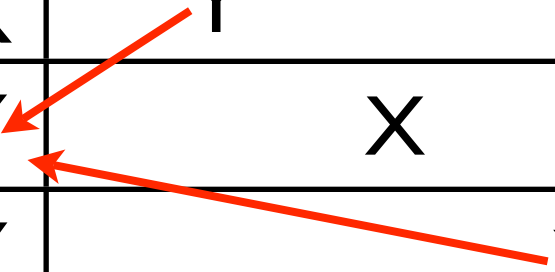
975-1040	X	Y	Z
980-1055	Y	X	
997-1072	Z		Y
1205-1278	Y	...	
		...	



The diagram illustrates the disambiguation of the letter 'Y' across different time periods. Red arrows point from the 'Y' entries in the first four rows to the 'Y' entry in the fifth row, indicating a mapping or relationship.

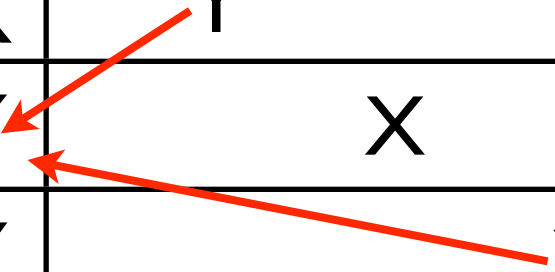
Disambiguation

975-1040	X	Y	Z
980-1055	Y	X	
997-1072	Z		Y
1205-1278	Y	...	
		...	



Disambiguation

975-1040	X	Y	Z
980-1055	Y	X	
997-1072	Z		Y
1205-1278	Y	...	
		...	



Might be useful to
look at when
people lived!

Collecting Time Information

- Birth and death years are incomplete
- Only 5,000 out of 24,000
- Examples:
 - 王翥 ??? 字一飛，成都人。篤學尚氣。吳曦謀反來請，翥陽病風瘡，潛往安丙，謀誅曦。事定，匿巴中，為農終身。
 - 王鈇 (? ~ 1149)，字承可，號亦樂居士，豫章人。秦檜舅氏王本觀復之子。檜薦於朝，除樞屬，嘗提舉浙東茶鹽，官至戶部侍郎。紹興十七年知廣州，十九年卒於官。有亦樂居士集。

Collecting Time Information

- But we can infer them from the text!
 - Years of death
“... died during the reign of ...”
 - Years of important events
“... was a presented scholar during the reign of ...”
 - Temple names of emperors
“... was summoned to the court by ...”
- Estimate when people lived

Pattern Search

- Searching for death dates, presented scholar dates etc.
- Cannot do simple search anymore
- Many “patterns” or “templates”

Pattern Search

- We want to find

元祐三年卒

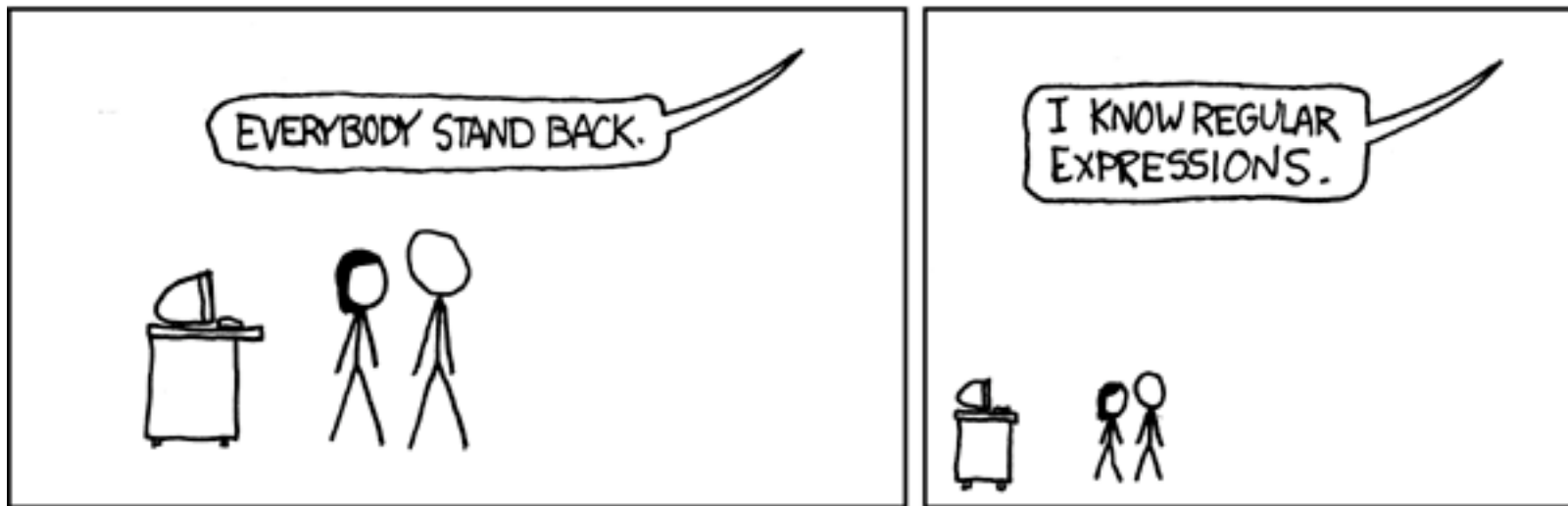
(died during the 3rd year of the reign of 元祐)

- The “search pattern” is

reign title number & year character death character

Regular Expressions

- Computational tools to “search for textual patterns”



Regular Expressions

- Using regular expressions to search for an email address
- ANY email address (without knowing)
- Search pattern:
username @ domain name .com
- Regular expression:
[a-z]+@[a-z]+\.

Regular Expressions

Regular Expressions

- Regular expression:
`[a-z]+@[a-z]+\.`com

Regular Expressions

- Regular expression:
`[a-z]+@[a-z]+\.`com

Elif Yamangil's Resume

email: elifyamangil@gmail.com

phone: (617) 909-6398

Education:

Harvard University, PhD candidate in
computer science, 2006 - today

Sabanci University, BS in computer
science, minor in Mathematics,
2001-2006

...

Regular Expressions

- Regular expression:
`[a-z]+@[a-z]+\.``com`
- Search my resume
using this regular
expression

Elif Yamangil's Resume

email: elifyamangil@gmail.com

phone: (617) 909-6398

Education:

Harvard University, PhD candidate in
computer science, 2006 - today

Sabanci University, BS in computer
science, minor in Mathematics,
2001-2006

...

Regular Expressions

- Regular expression:
[a-z]+@[a-z]+\.com
- Search my resume
using this regular
expression
- Finds my email
address!

Elif Yamangil's Resume

email: elifyamangil@gmail.com

phone: (617) 909-6398

Education:

Harvard University, PhD candidate in
computer science, 2006 - today

Sabanci University, BS in computer
science, minor in Mathematics,
2001-2006

...

Regular Expressions

- Regular expression:
`[a-z]+@[a-z]+\.``com`
- Search my resume
using this regular
expression
- Finds my email
address!

Elif Yamangil's Resume

email: `elifyamangil@gmail.com`

phone: (617) 909-6398

Education:

Harvard University, PhD candidate in
computer science, 2006 - today

Sabanci University, BS in computer
science, minor in Mathematics,
2001-2006

...

Regular Expressions

- Same idea!
- First convert Chinese numbers to Arabic
- Then use regular expression:

元祐[0-9]+年卒

Regular Expressions

- Of course we want to search for any reign title:

(元祐|元豐|...|紹興)[0-9]+年卒

- And use various other patterns:

(元祐|元豐|...|紹興)[0-9]+年[0-9]+月卒

(元祐|元豐|...|紹興)[0-9]+年[0-9]+月[0-9]+日卒

(元祐|元豐|...|紹興)(初|中|末|間)卒

(元祐|元豐|...|紹興) [0-9]+年(春|夏|秋|冬)卒

Regular Expressions

Regular Expressions

Using the regular expression:

(元祐|元豐|...|紹興)[0-9]+年卒

Regular Expressions

Using the regular expression:

(元祐|元豐|...|紹興)[0-9]+年卒

WDY
丁罕 (? ~999), 潁州人。應募補衛士, 以戰功累遷指揮使。淳化中為澤州團練使, 知霸州, 河決, 以私錢募築, 民咸德之, 擢領靈環路都部署, 破李繼遷有殊功。後拜密州觀察使, 徙貝州。咸平二年卒, 子守德能世其家。
丁明 (1127~1211), 舊名騫, 字希閔, 後改名明, 字子公, 金壇人, 權子。閉門讀書二十年, 手編事類及史通考等書百餘卷。奉祠家居, 嘉定四年卒, 年八十五。鄉里私謚博雅先生。

Regular Expressions

Using the regular expression:

(元祐|元豐|...|紹興)[0-9]+年卒

WDY

丁罕 (? ~999), 潁州人。應募補衛士, 以戰功累遷指揮使。淳化中為澤州團練使, 知霸州, 河決, 以私錢募築, 民咸德之, 擢領靈環路都部署, 破李繼遷有殊功。後拜密州觀察使, 徙貝州。
咸平二年卒, 子守德能世其家。

丁明 (1127~1211), 舊名騫, 字希閔, 後改名明, 字子公, 金壇人, 權子。閉門讀書二十年, 手編事類及史通考等書百餘卷。奉祠家居, 嘉定四年卒, 年八十五。鄉里私謚博雅先生。

Regular Expressions

Using the regular expression:

(元祐|元豐|...|紹興)[0-9]+年卒

WDY

丁罕 (? ~999), 潁州人。應募補衛士, 以戰功累遷指揮使。淳化中為澤州團練使, 知霸州, 河決, 以私錢募築, 民咸德之, 擢領靈環路都部署, 破李繼遷有殊功。後拜密州觀察使, 徙貝州。
咸平二年卒, 子守德能世其家。

丁明 (1127~1211), 舊名騫, 字希閔, 後改名明, 字子公, 金壇人, 權子。閉門讀書二十年, 手編事類及史通考等書百餘卷。奉祠家居, 嘉定四年卒, 年八十五。鄉里私謚博雅先生。

Regular Expressions

- Presented Scholar degree date patterns

(元祐|元豐|...|紹興)[0-9]+年[四甲第十八名]{,9}進士
(元祐|元豐|...|紹興) (初|中|末|間)[四甲第十八名]{,9}進士

- Temple names can be handled using simple search

Collected Dates

Collected Dates

ID: 24603, temple name (mid year: 1093)

ID: 24610, temple name (mid year: 1010)

ID: 24615, temple name (mid year: 1244)

ID: 24619, death date (year: 973)

ID: 24646, death date (year: 1030, month: 6, day: 15)

ID: 24670, temple name (mid year: 1065)

ID: 24695, death date (year: 1082, month: 10)

ID: 24708, death date (year: 1082, season: 4)

ID: 24709, presented scholar (year: 1106)

ID: 24710, presented scholar (year: 1053)

ID: 24720, presented scholar (year: 1148)

ID: 24723, temple name (mid year: 1076)

Collected Dates

ID: 24603, temple name (mid year: 1093)

ID: 24610, temple name (mid year: 1010)

ID: 24615, temple name (mid year: 1244)

ID: 24619, death date (year: 973)

ID: 24646, death date (year: 1030, month: 6, day: 15)

ID: 24670, temple name (mid year: 1065)

ID: 24695, death date (year: 1082, month: 10)

ID: 24708, death date (year: 1082, season: 4)

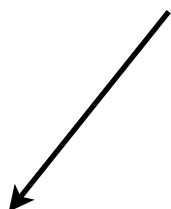
ID: 24709, presented scholar (year: 1106)

ID: 24710, presented scholar (year: 1053)

ID: 24720, presented scholar (year: 1148)

ID: 24723, temple name (mid year: 1076)

This person
died on 6-15-1030



Collected Dates

ID: 24603, temple name (mid year: 1093)

ID: 24610, temple name (mid year: 1010)

ID: 24615, temple name (mid year: 1244)

ID: 24619, death date (year: 973)

ID: 24646, death date (year: 1030, month: 6, day: 15)

ID: 24670, temple name (mid year: 1065)

ID: 24695, death date (year: 1082, month: 10)

ID: 24708, death date (year: 1082, season: 4)

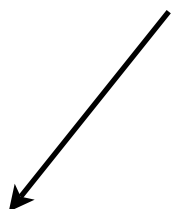
ID: 24709, presented scholar (year: 1106)

ID: 24710, presented scholar (year: 1053)

ID: 24720, presented scholar (year: 1148)

ID: 24723, temple name (mid year: 1076)

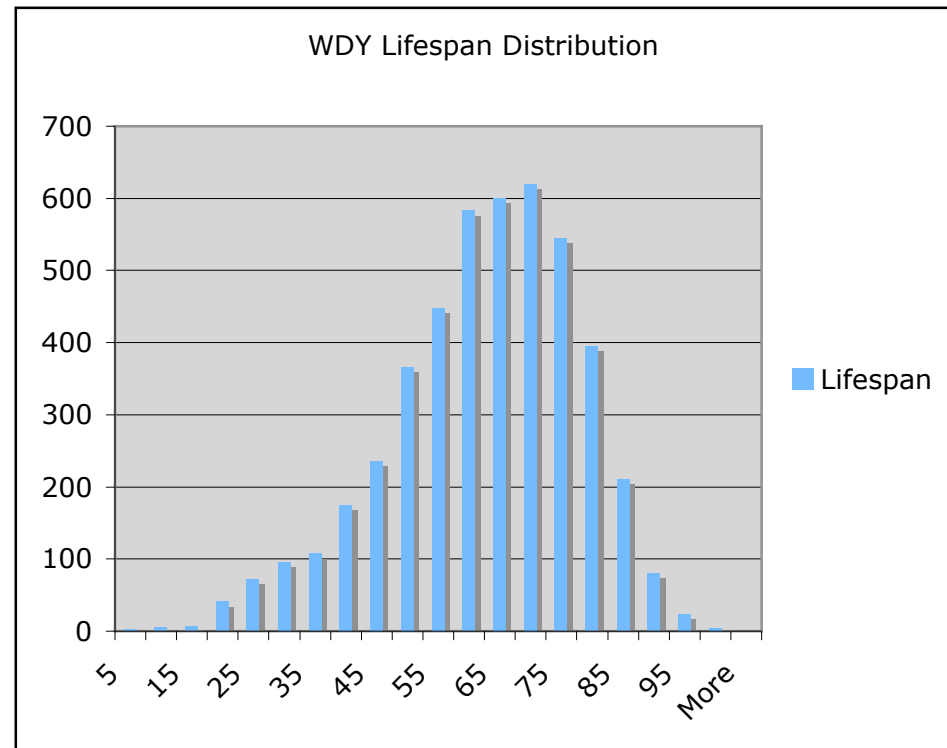
This person
died on 6-15-1030



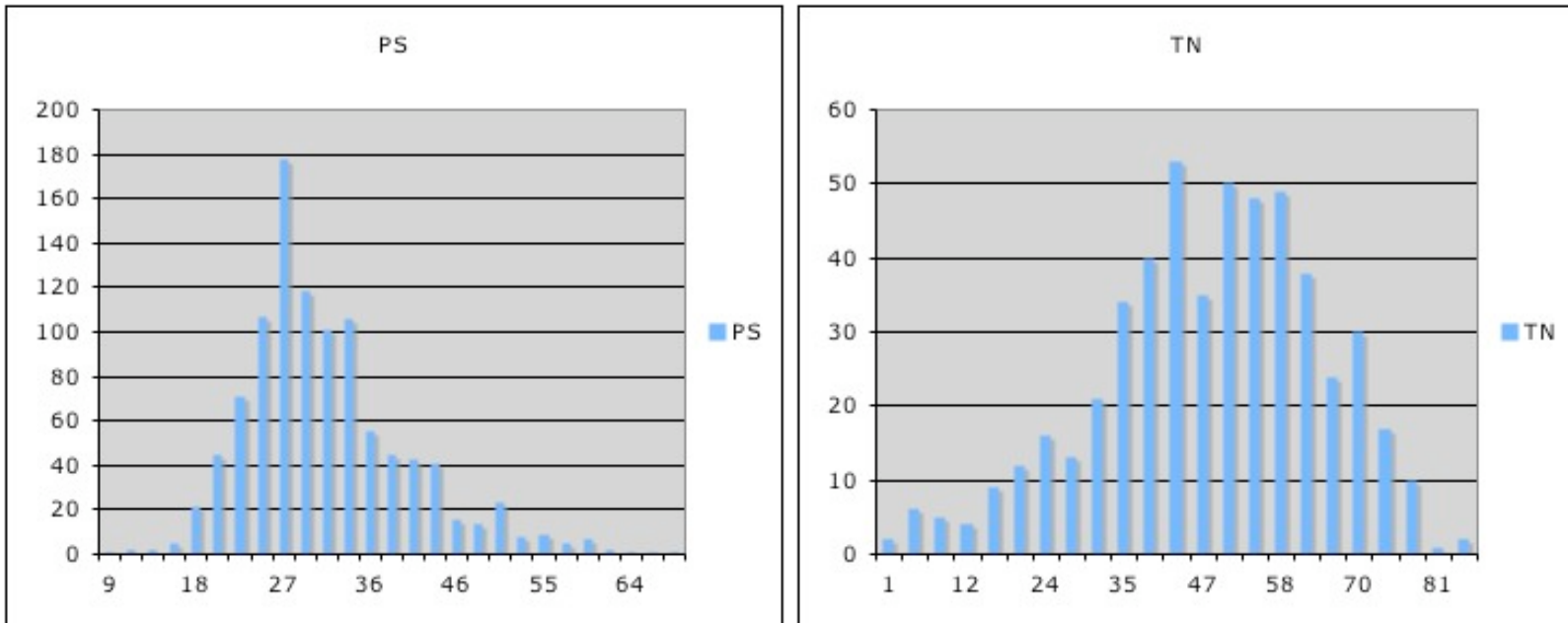
91% of the death
dates found agreed
on year with
Arabic death years
(9% must be
incorrect!)

Collected Dates

- Lifespan distribution
- Average lifetime ≈ 60 years



Collected Dates



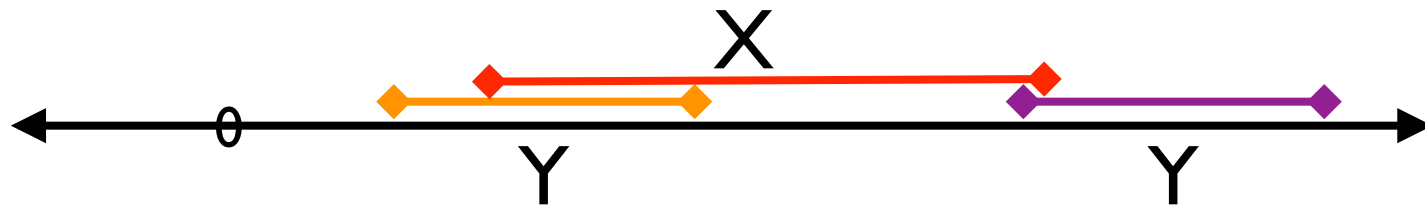
- PS corresponds to $\approx 30^{\text{th}}$ year of life
- TN correspond to $\approx 45^{\text{th}}$ year of life

Reconstructing Lifetimes

- We “re-construct” unknown lifespans
- How do we use these dates?
 - Death year = 60th year of life
 - PS year = 30th year of life
 - TN midpoint = 45th year of life
- In this order
- 19,000 unknown → 14,000 unknown
- 2,000 “other” dates were found

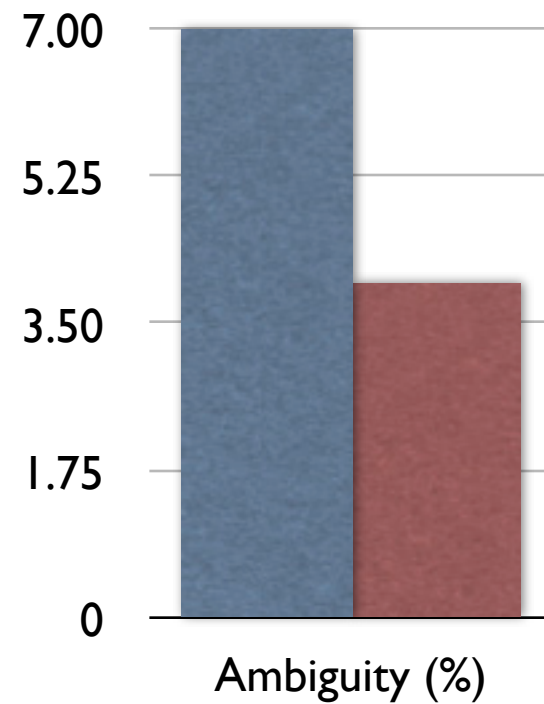
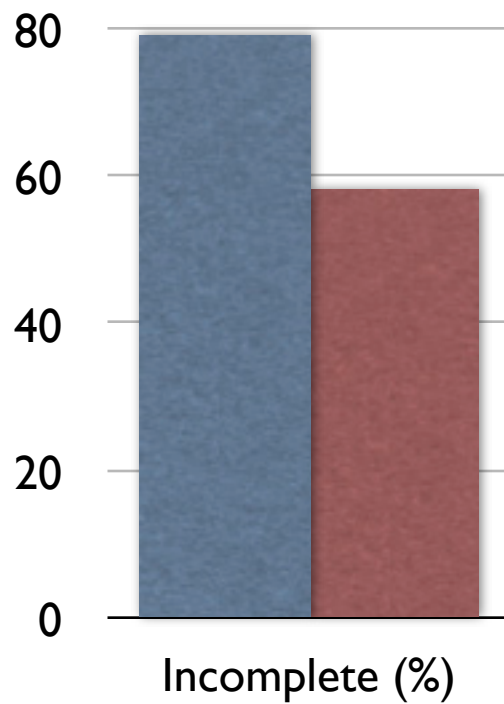
Disambiguating Name Co-occurrences

- How do we use lifespans to disambiguate?

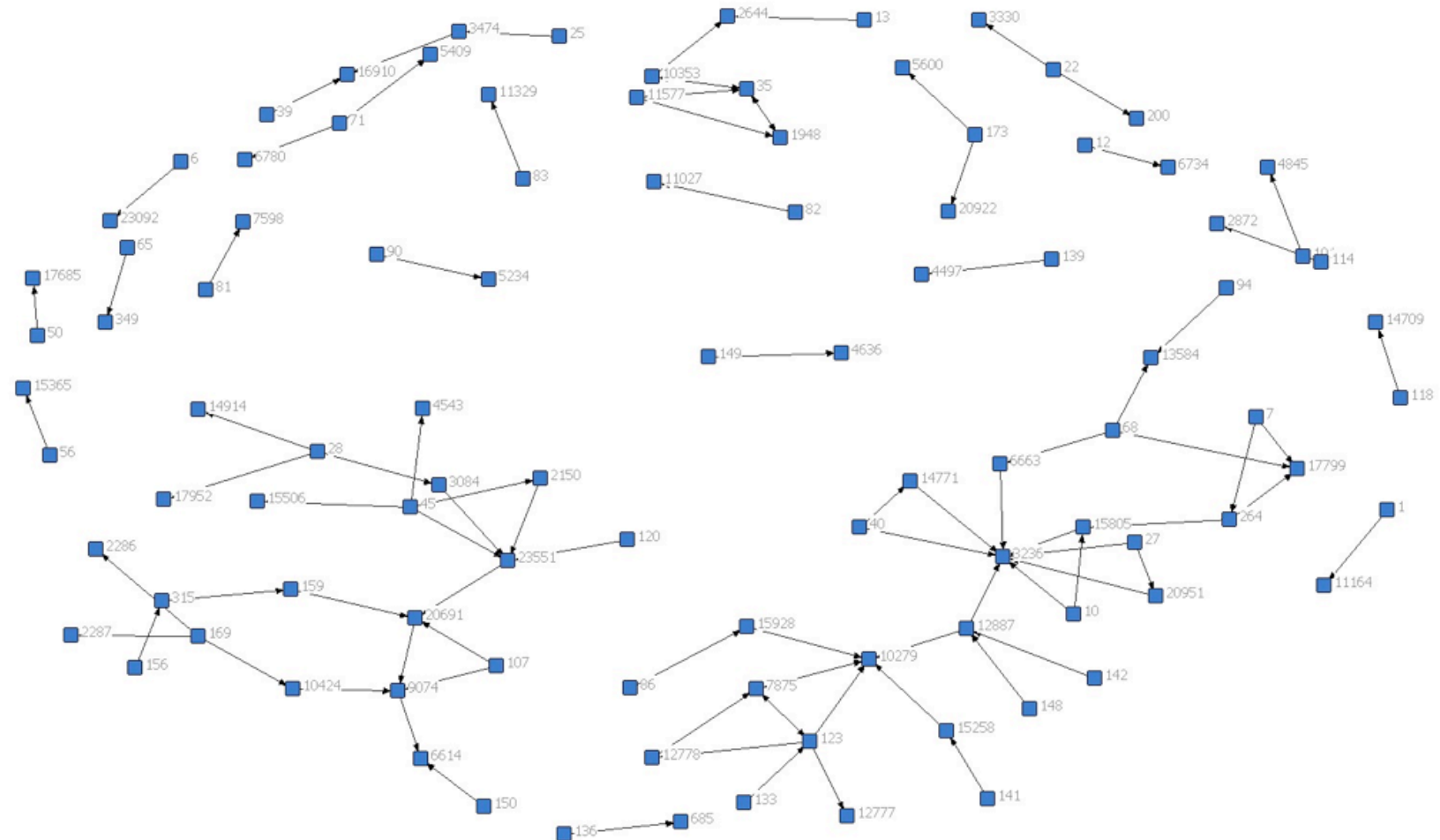


- We check for longest overlap

Results

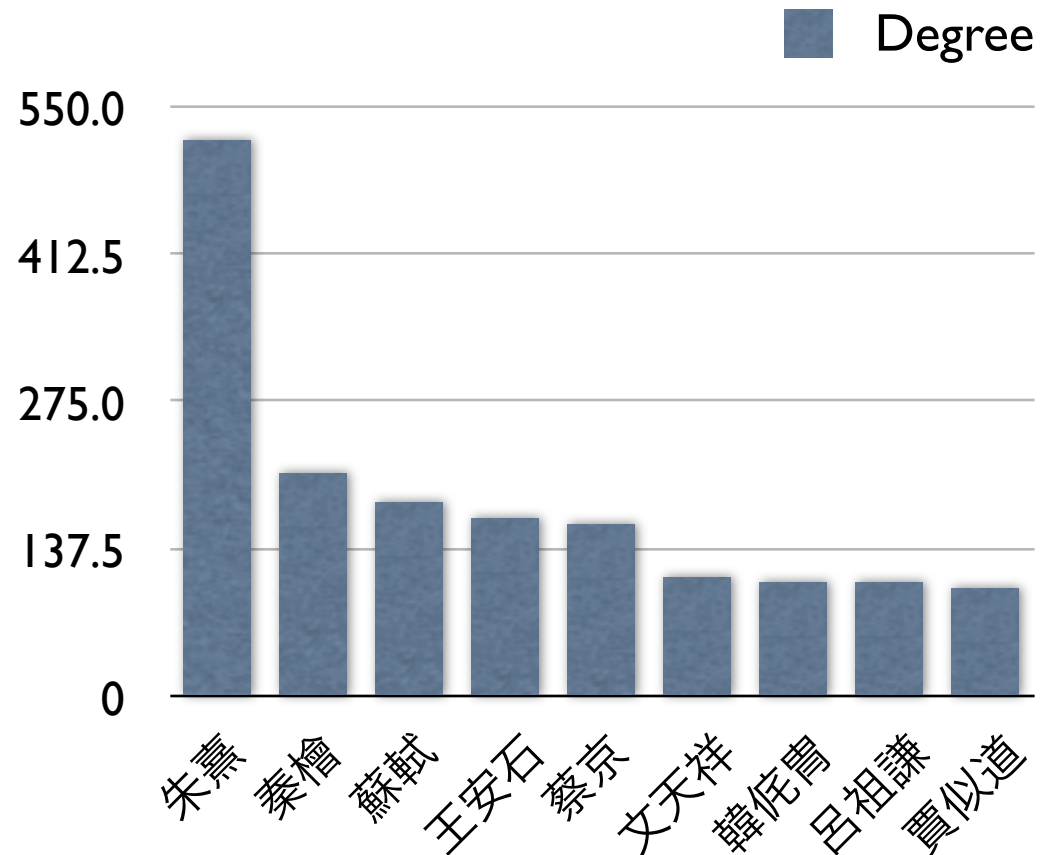


Graph Analysis

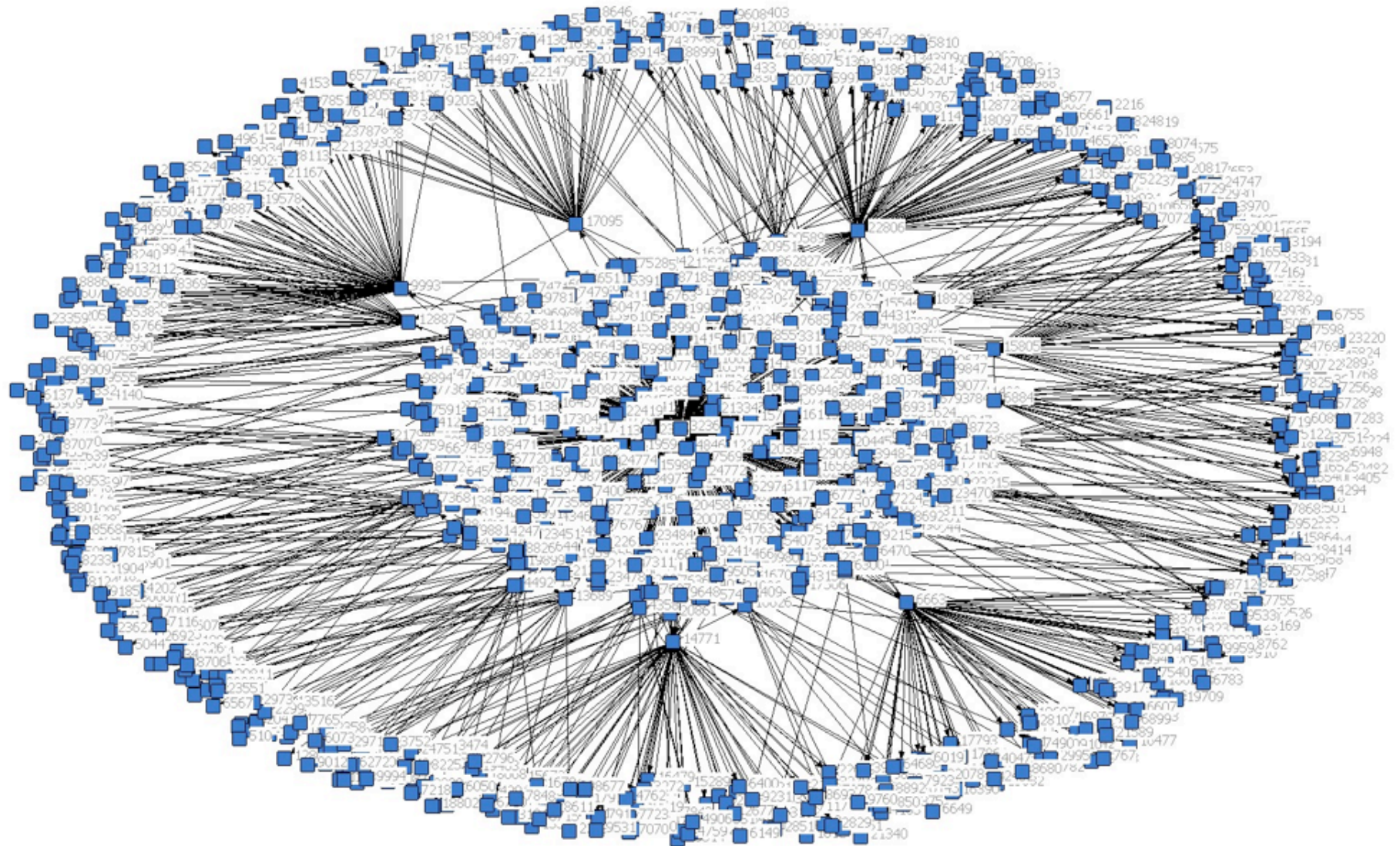


Graph Analysis

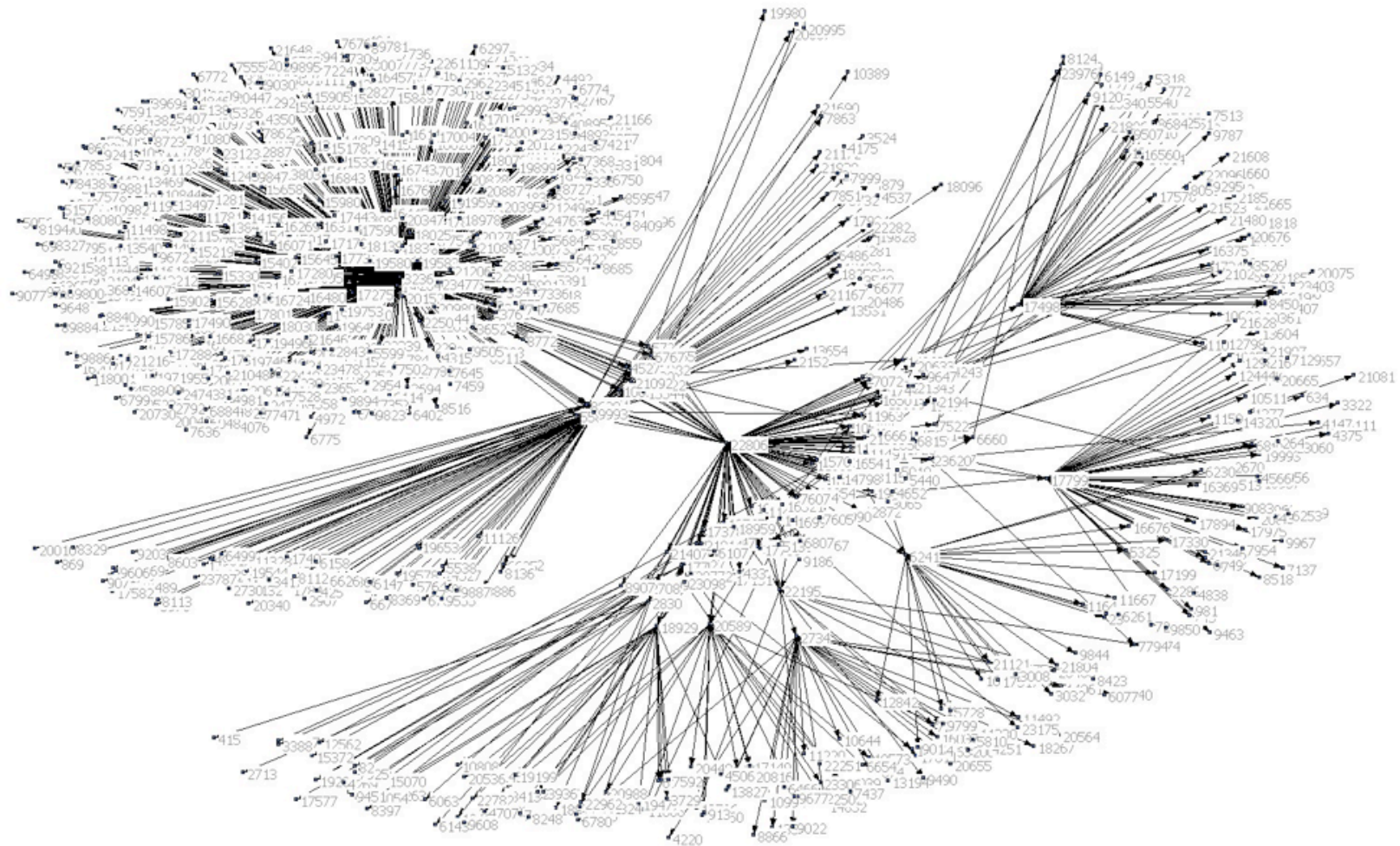
- Degree distribution
- Important figures have high degree



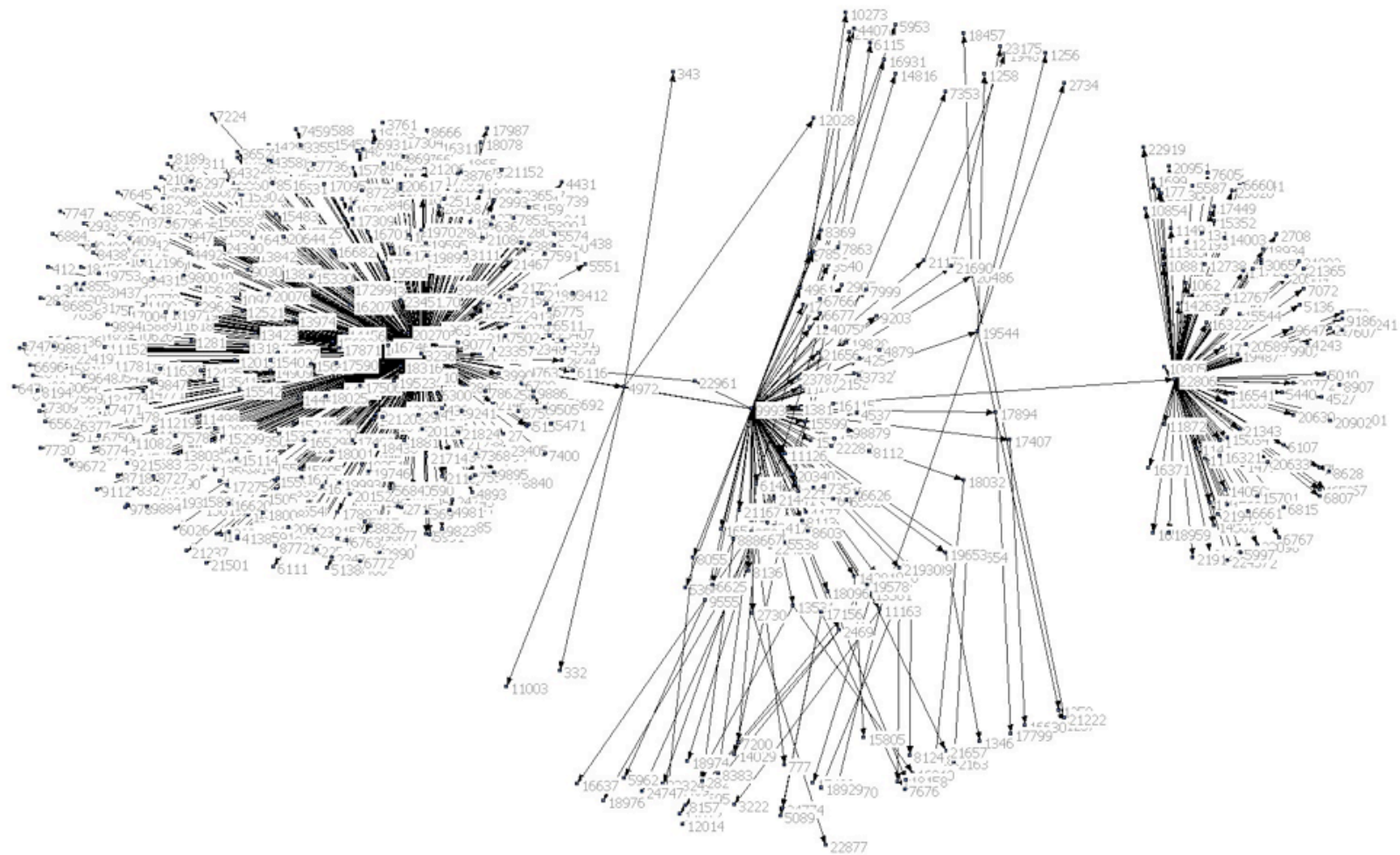
Graph Analysis



Graph Analysis



Graph Analysis



Summary

- Searching for **name co-occurrences**
- Regular expressions help us pick up **important dates**
- From which we infer unknown **lifespans**
- Lifespans help us **disambiguate**
- Graphs hint at **interesting structure**

Next Steps

- Evaluate disambiguation (using hand-annotation)
- Improve search and disambiguation
 - Kinship information
 - Place names
 - Alternate names (capping names, posthumous names, studio names)
- Collecting data is interesting in its own right

Thanks!