PANEL 300. **C** **Saturday** 5:00pm-7:00pm Gregory A/B, 2nd Level

WORKSHOP:

Computational Tools and Digital Resources for Chinese History and Literature

Michael A. Fuller
University of California, Irvine

From Texts to Databases:

The Computer Modeling and Analysis of Historical and Literary Materials

First Question: For what sort of historical or literary material do you want to develop a model?
Next Question: Why? What do you want to do with the data?

For example, let us consider letters written from one person to another. There are many reasons to study letters, some more amenable to computer analysis than others.

- We could be interested in **content**:
  - What are the topics?
  - What people are mentioned
  - What places or events or texts are mentioned
- We could be interested in **stylistic and rhetorical issues**:
  - What levels of language are used?
  - Are there particular repeated persuasive strategies?
  - How do the authors present themselves and their recipients?

- We also could be interested in the **sociology of production and circulation of letters**:
  - Are there discrete networks of people writing one another?
  - Do these networks have particular characteristics in terms of geographic distribution or social status?

How we model a "letter" depends on what we want to look at.
BUT: a prudent design preserves "hooks" so that we do not need to repeat work later.
This requires thought and planning.

Basic Entity: **LETTER**:

Author (Person)
Recipients (List of People)
Date of composition
Place of composition
Text of Letter

For sociological analysis of a group of letters, we might want to know various facts about the people involved in the letter: place of origin, status, age, birth date, and so on.

Because we do not want to repeat that information for authors and recipients in the record for each letter, we create a separate entity:

PEOPLE:

Name
Place
Birth date
Status, etc.

We also notice that there are other categories of information for **LETTERS** and **PEOPLE** that are complex and in fact should be considered *entities*:

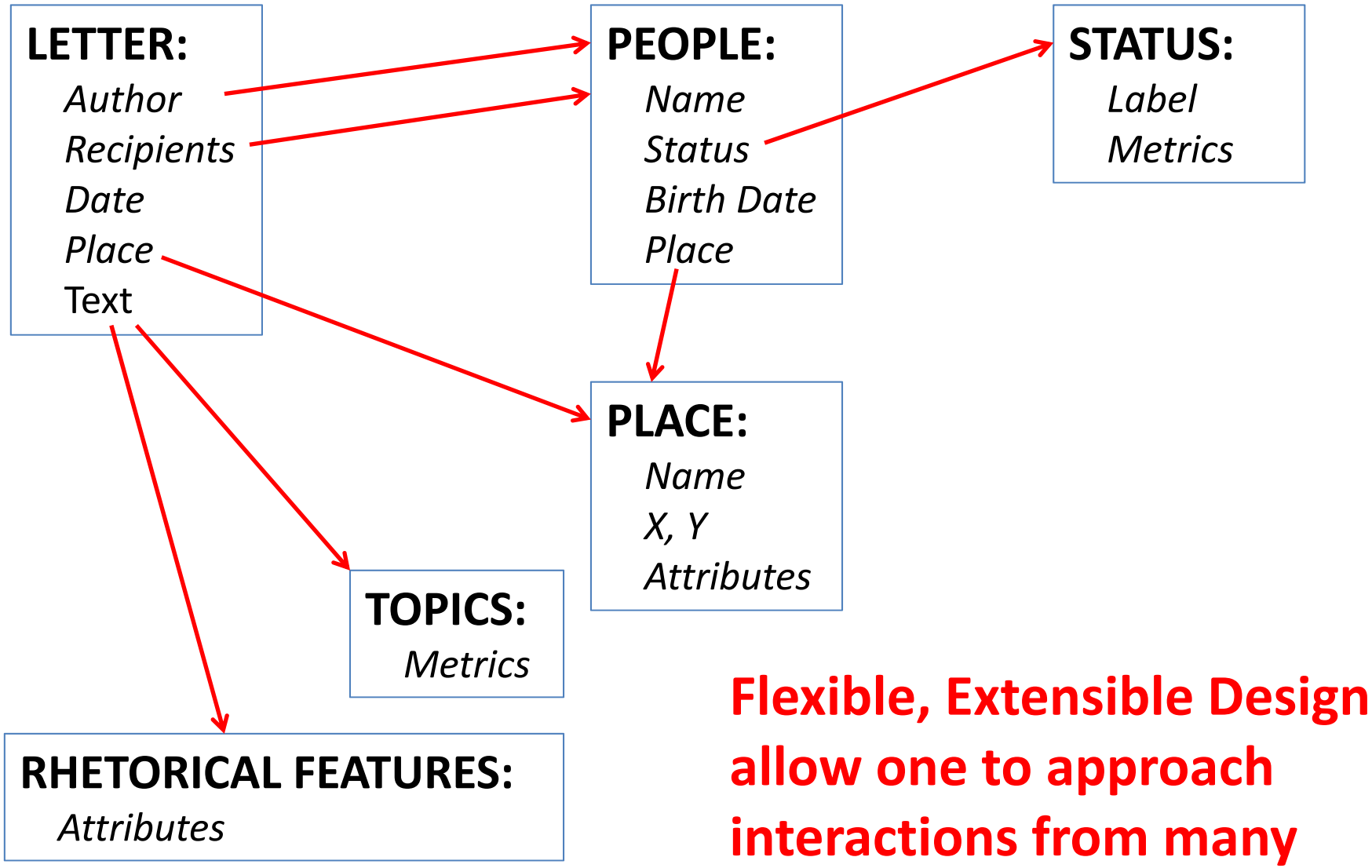
PLACES:

Name
Type (prefecture, county, etc.)
x-coordinate
Y-coordinate

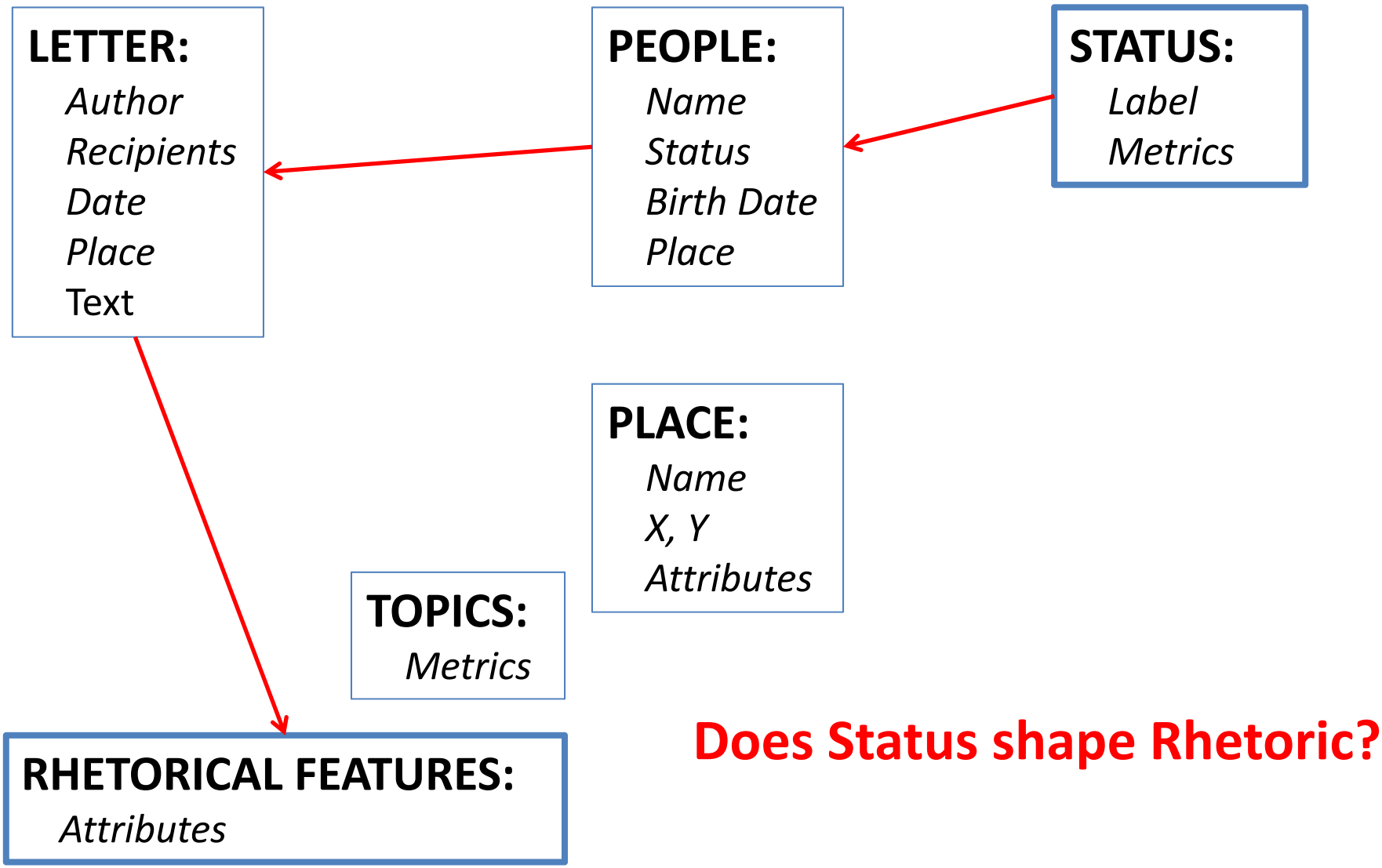
STATUS:

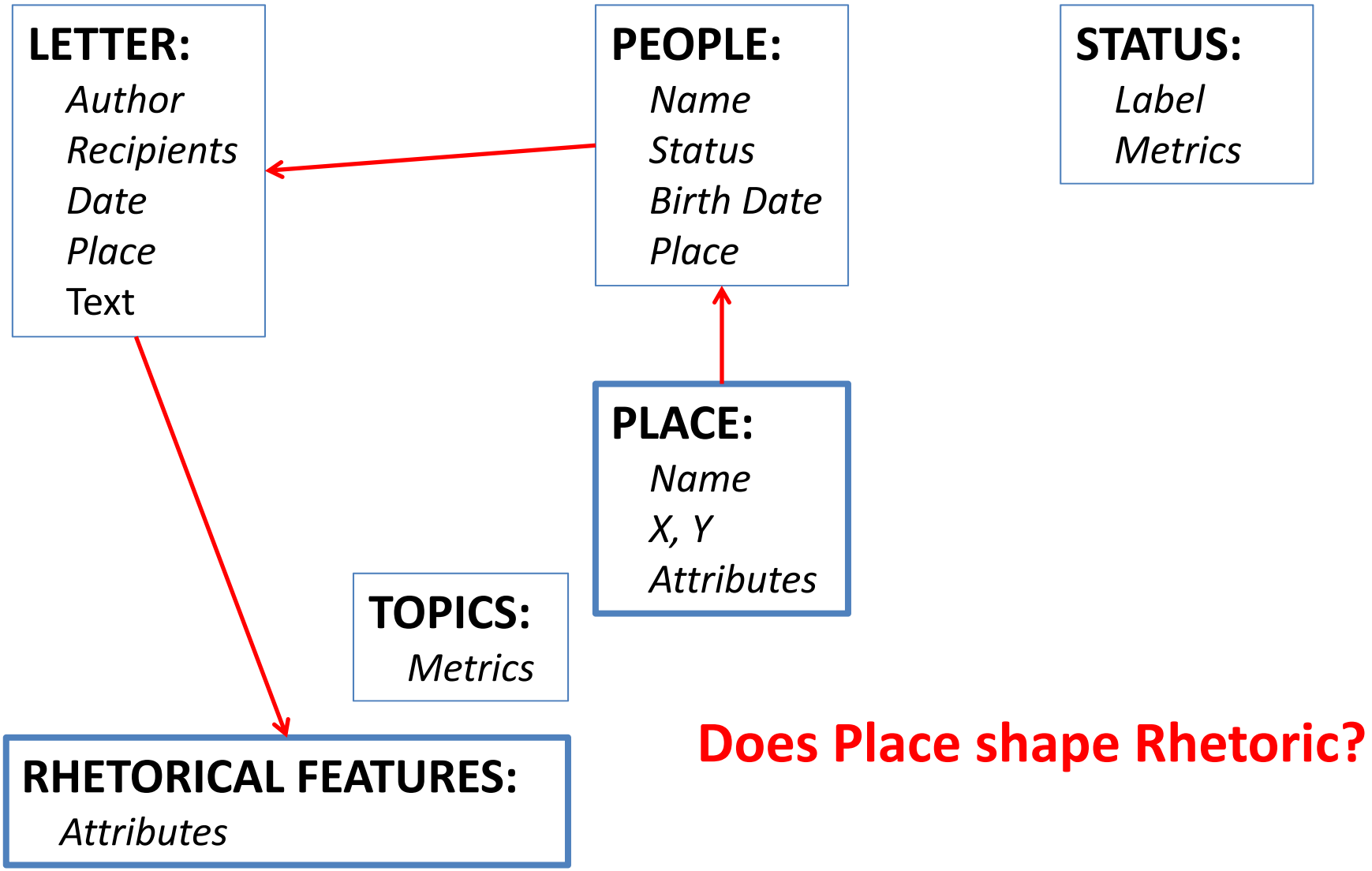
Label
Metrics for status (date of jinshi, official title, etc.)

Now we can think about how these various entities interact:



Flexible, Extensible Design allow one to approach interactions from many angles

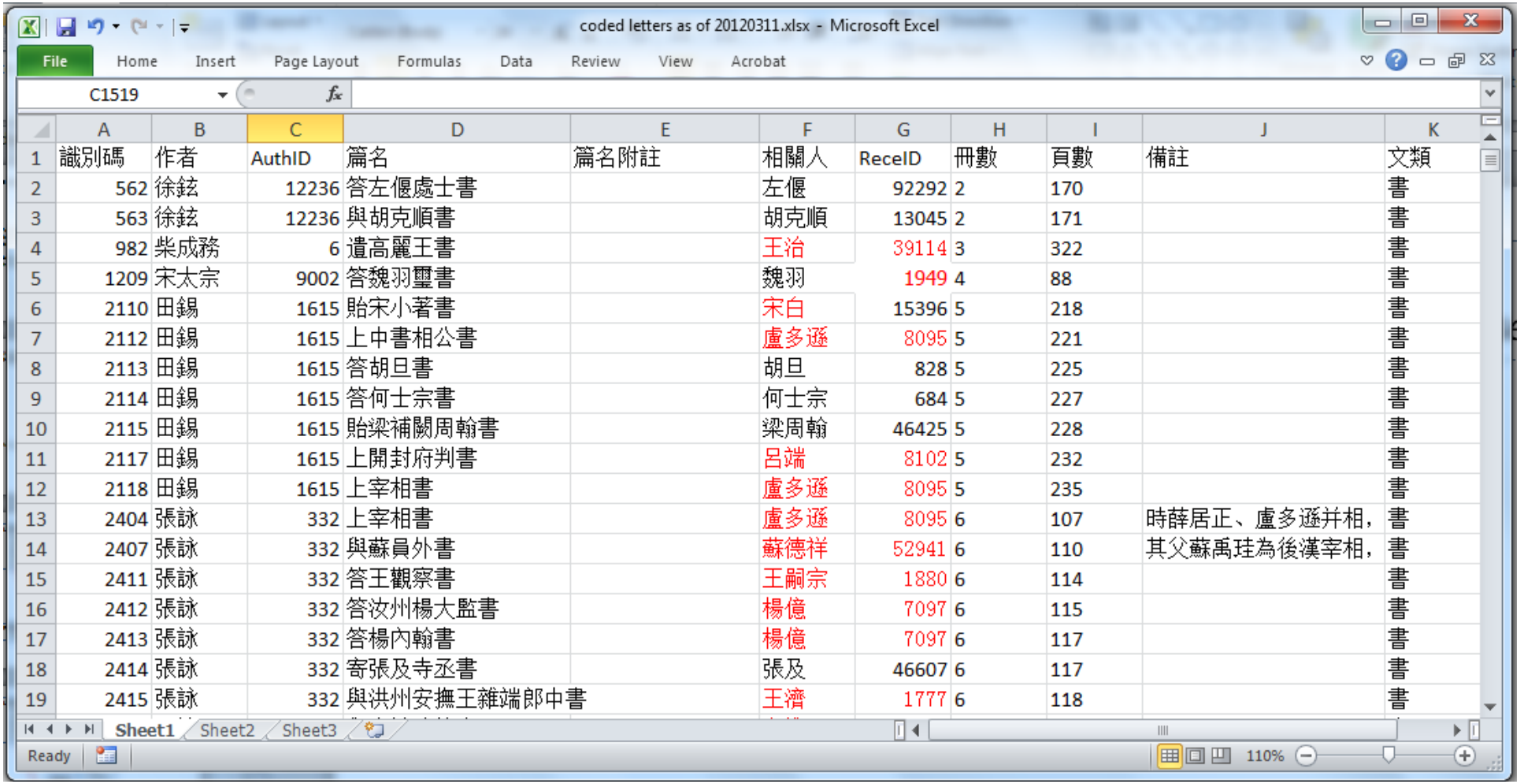

LETTER:
Author
Recipients
Date
Place
Text
PEOPLE:
Name
Status
Birth Date
Place
STATUS:
Label
Metrics
PLACE:
Name
X, Y
Attributes
TOPICS:
Metrics
RHETORICAL FEATURES:
Attributes
Does Status shape Rhetoric?


LETTER:
Author
Recipients
Date
Place
Text
PEOPLE:
Name
Status
Birth Date
Place
STATUS:
Label
Metrics
PLACE:
Name
X, Y
Attributes
TOPICS:
Metrics
RHETORICAL FEATURES:
Attributes
Does Place shape Rhetoric?

A Preliminary Model:

The 8,000 Letters (so far) in the *Complete Song Dynasty Prose*

It started as an Excel Spreadsheet of the Titles, Authors, Recipients:



coded letters as of 20120311.xlsx - Microsoft Excel

File Home Insert Page Layout Formulas Data Review View Acrobat

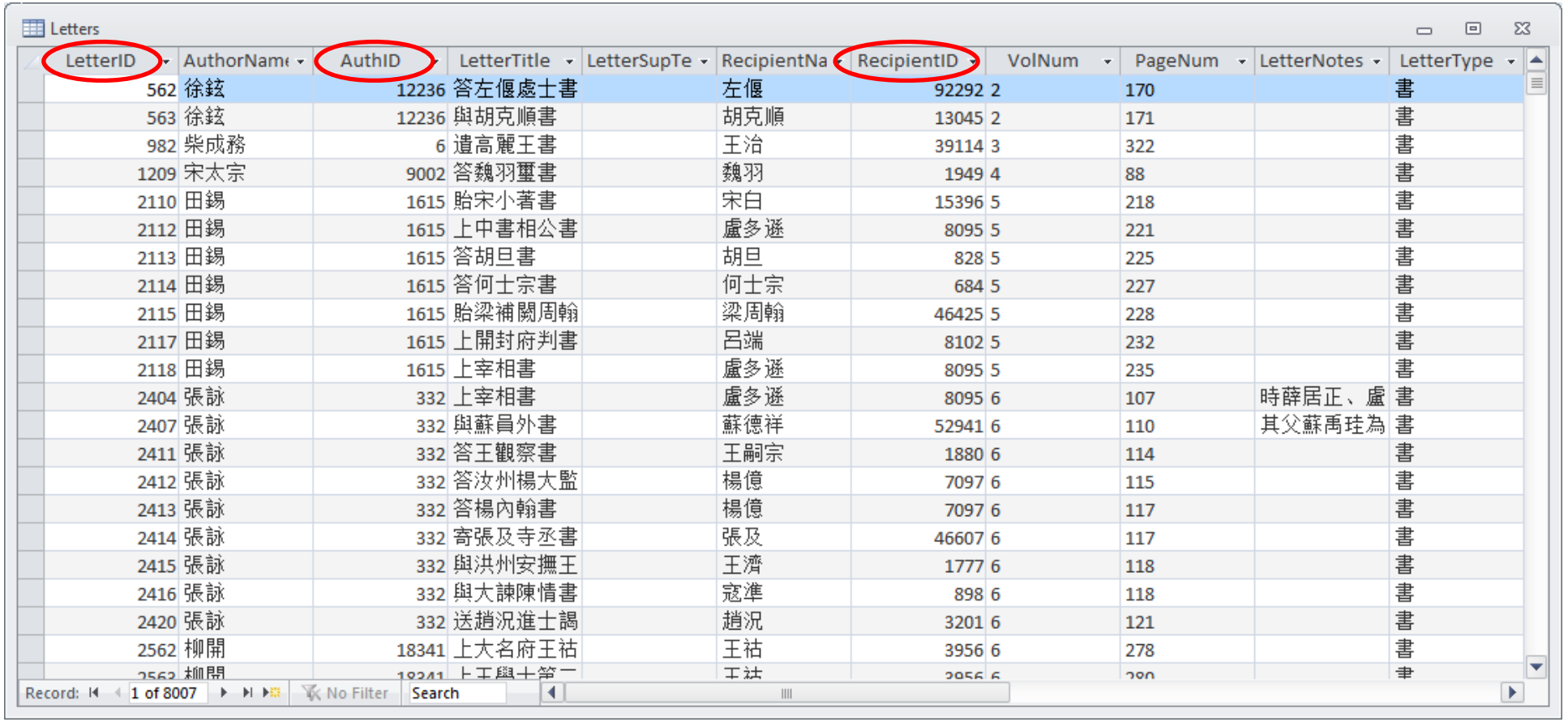
C1519

	A	B	C	D	E	F	G	H	I	J	K
1	識別碼	作者	AuthID	篇名	篇名附註	相關人	ReceID	冊數	頁數	備註	文類
2	562	徐鉉	12236	答左偃處士書		左偃	92292	2	170		書
3	563	徐鉉	12236	與胡克順書		胡克順	13045	2	171		書
4	982	柴成務	6	遣高麗王書		王治	39114	3	322		書
5	1209	宋太宗	9002	答魏羽璽書		魏羽	1949	4	88		書
6	2110	田錫	1615	貽宋小著書		宋白	15396	5	218		書
7	2112	田錫	1615	上中書相公書		盧多遜	8095	5	221		書
8	2113	田錫	1615	答胡旦書		胡旦	828	5	225		書
9	2114	田錫	1615	答何士宗書		何士宗	684	5	227		書
10	2115	田錫	1615	貽梁補闕周翰書		梁周翰	46425	5	228		書
11	2117	田錫	1615	上開封府判書		呂端	8102	5	232		書
12	2118	田錫	1615	上宰相書		盧多遜	8095	5	235		書
13	2404	張詠	332	上宰相書		盧多遜	8095	6	107	時薛居正、盧多遜并相，	書
14	2407	張詠	332	與蘇員外書		蘇德祥	52941	6	110	其父蘇禹珪為後漢宰相，	書
15	2411	張詠	332	答王觀察書		王嗣宗	1880	6	114		書
16	2412	張詠	332	答汝州楊大監書		楊億	7097	6	115		書
17	2413	張詠	332	答楊內翰書		楊億	7097	6	117		書
18	2414	張詠	332	寄張及寺丞書		張及	46607	6	117		書
19	2415	張詠	332	與洪州安撫王雜端郎中書		王濟	1777	6	118		書

Sheet1 Sheet2 Sheet3

Ready 110%

The spreadsheet, imported into Access, becomes a table:



Letters

LetterID	AuthorName	AuthID	LetterTitle	LetterSupTe	RecipientNa	RecipientID	VolNum	PageNum	LetterNotes	LetterType
562	徐鉉	12236	答左偃處士書		左偃	92292	2	170		書
563	徐鉉	12236	與胡克順書		胡克順	13045	2	171		書
982	柴成務	6	遺高麗王書		王治	39114	3	322		書
1209	宋太宗	9002	答魏羽璽書		魏羽	1949	4	88		書
2110	田錫	1615	貽宋小著書		宋白	15396	5	218		書
2112	田錫	1615	上中書相公書		盧多遜	8095	5	221		書
2113	田錫	1615	答胡旦書		胡旦	828	5	225		書
2114	田錫	1615	答何士宗書		何士宗	684	5	227		書
2115	田錫	1615	貽梁補闕周翰		梁周翰	46425	5	228		書
2117	田錫	1615	上開封府判書		呂端	8102	5	232		書
2118	田錫	1615	上宰相書		盧多遜	8095	5	235		書
2404	張詠	332	上宰相書		盧多遜	8095	6	107	時薛居正、盧	書
2407	張詠	332	與蘇員外書		蘇德祥	52941	6	110	其父蘇禹珪為	書
2411	張詠	332	答王觀察書		王嗣宗	1880	6	114		書
2412	張詠	332	答汝州楊大監		楊億	7097	6	115		書
2413	張詠	332	答楊內翰書		楊億	7097	6	117		書
2414	張詠	332	寄張及寺丞書		張及	46607	6	117		書
2415	張詠	332	與洪州安撫王		王濟	1777	6	118		書
2416	張詠	332	與大諫陳情書		寇準	898	6	118		書
2420	張詠	332	送趙況進士謁		趙況	3201	6	121		書
2562	柳開	18341	上大名府王祐		王祐	3956	6	278		書

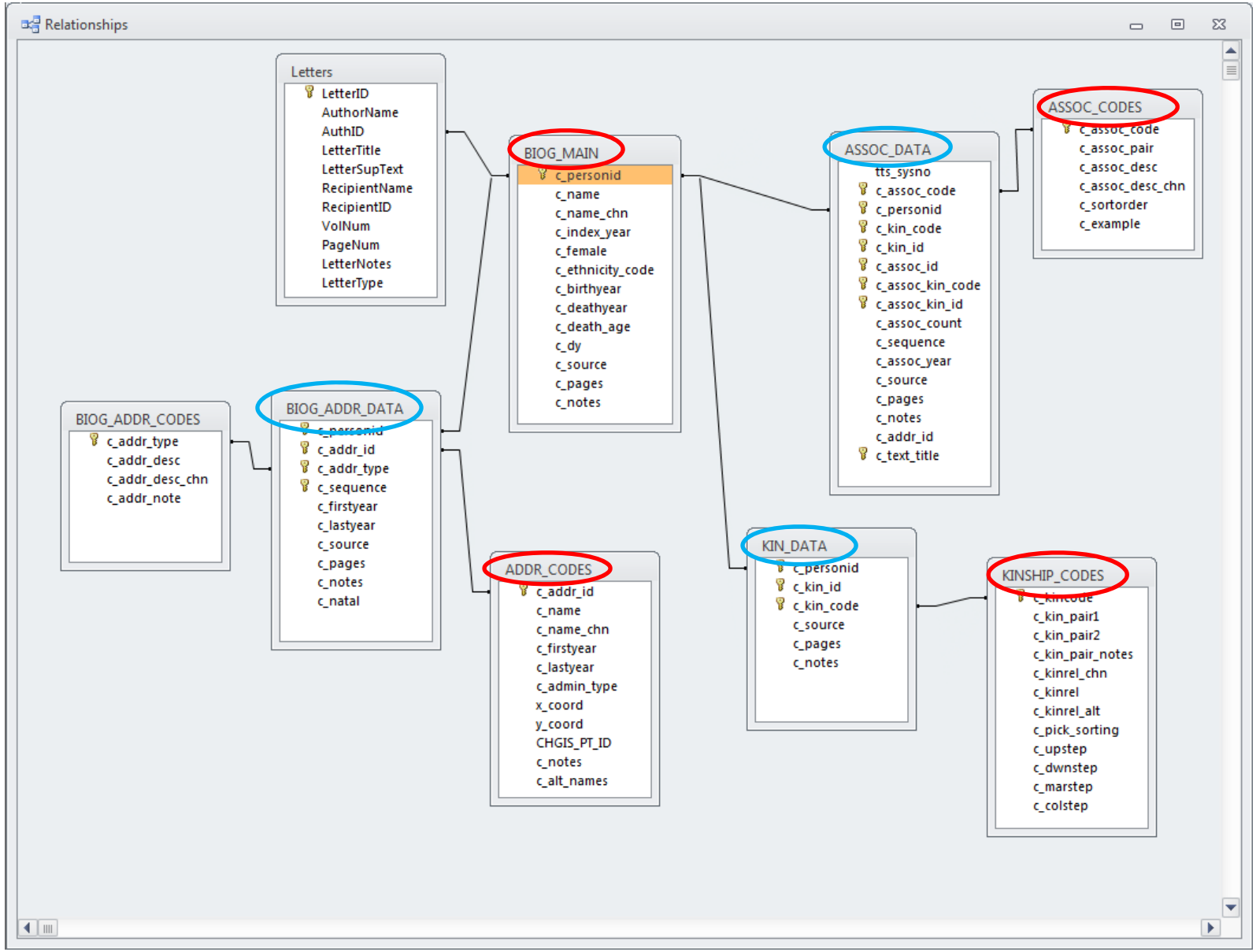
Record: 1 of 8007 | No Filter | Search

Note the structure of the data:

The letters have IDs

The Authors and Recipients have IDs: these are the CBDB IDs

Because the people use CBDB IDs, I can import CBDB tables for such entities as: PEOPLE, PLACE , KINSHIP, and ASSOCIATION (and can import more if needed later)



However, starting from a spreadsheet can lead to bad design because one tries to model the data as single records in a flat table.

The spreadsheet I was given, for example, had only one column for the recipient.
What happened when a letter had more that one addressee?
They created a new record (and a new letter ID) for each recipient:

Letters

LetterID	AuthorName	AuthID	LetterTitle	LetterSupText	RecipientName	RecipientID	VolNum	PageNum	LetterNotes	LetterType
51652	黃庭堅	7111	與洪氏四甥書	一	洪羽	10618	105	108		書
51653	黃庭堅	7111	與洪氏四甥書	二	洪炎	10616	105	108		書
51654	黃庭堅	7111	與洪氏四甥書	三	洪朋	10615	105	109		書
51655	黃庭堅	7111	與洪氏四甥書	四	洪芻	10617	105	109		書

“Recipients” is a classic **one-to-many** relationship: **one** letter may have **many** addressees. To allow us to record these addressees, it is best to create a separate table:

RECIPIENTS

Letter ID
Recipient ID
Recipient Place
Receipt Date
etc.

Note that there is no point repeating any information about the letter or about the recipient in this table: we have that information elsewhere in the database.

The principle of not repeating information—that is, recording it just once in the database—is called **normalization**.

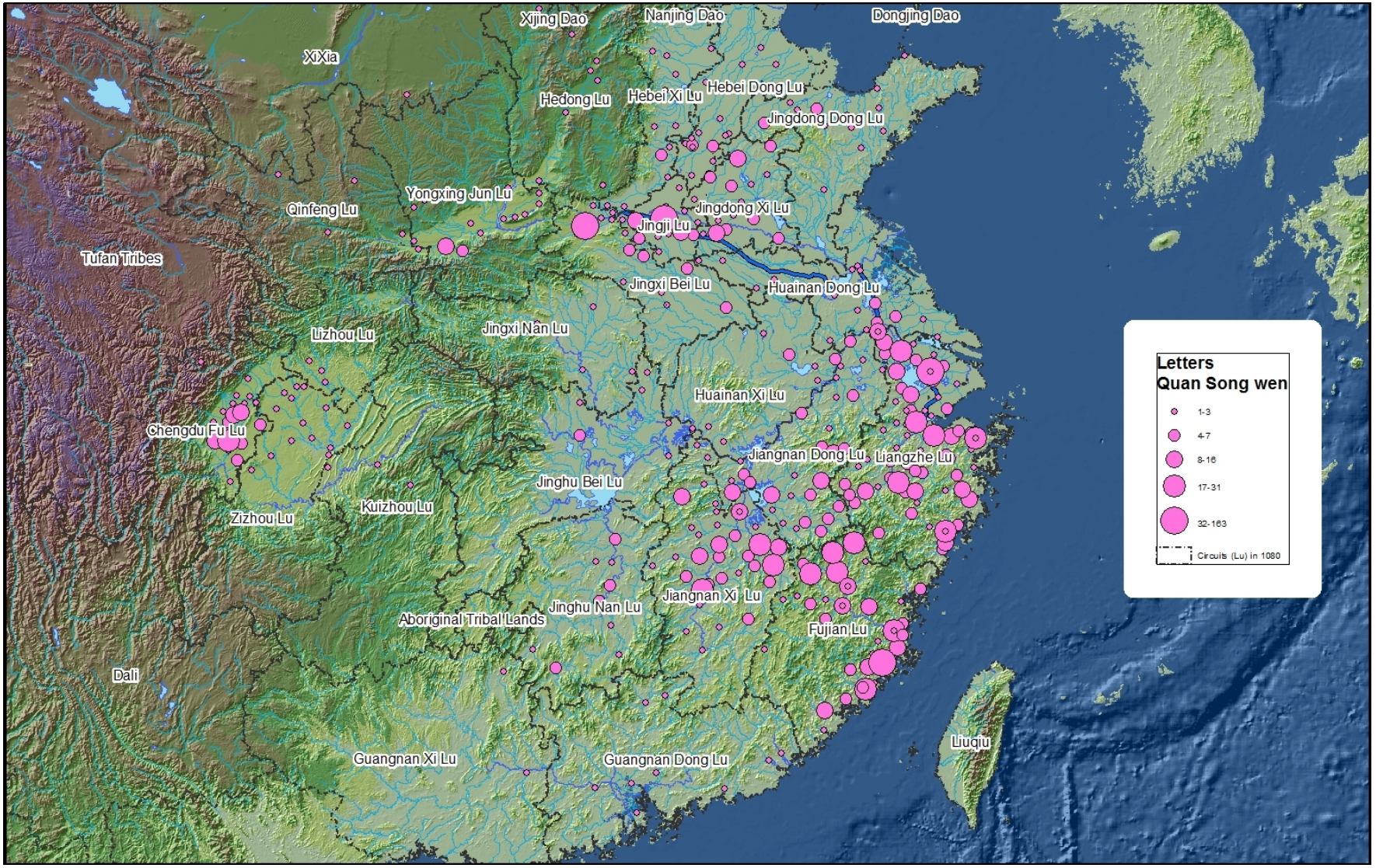
(You’ll note, however, that the LETTERS table violates this principle in recording people’s names.)

What can we do with the data at present?

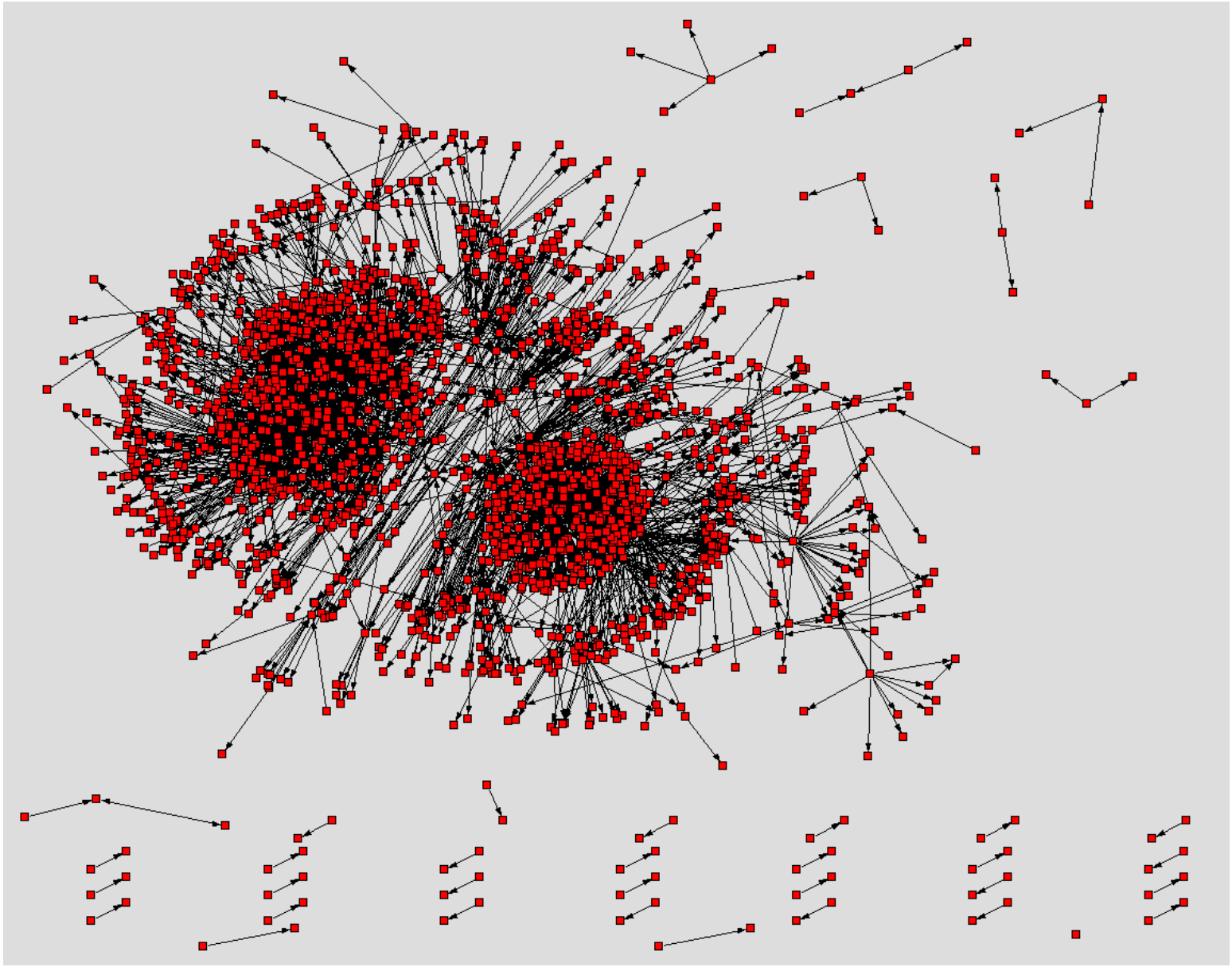
First, we can sort out the data a bit:

Among the 8,004 letters, there are 2,793 sets of authors and recipients

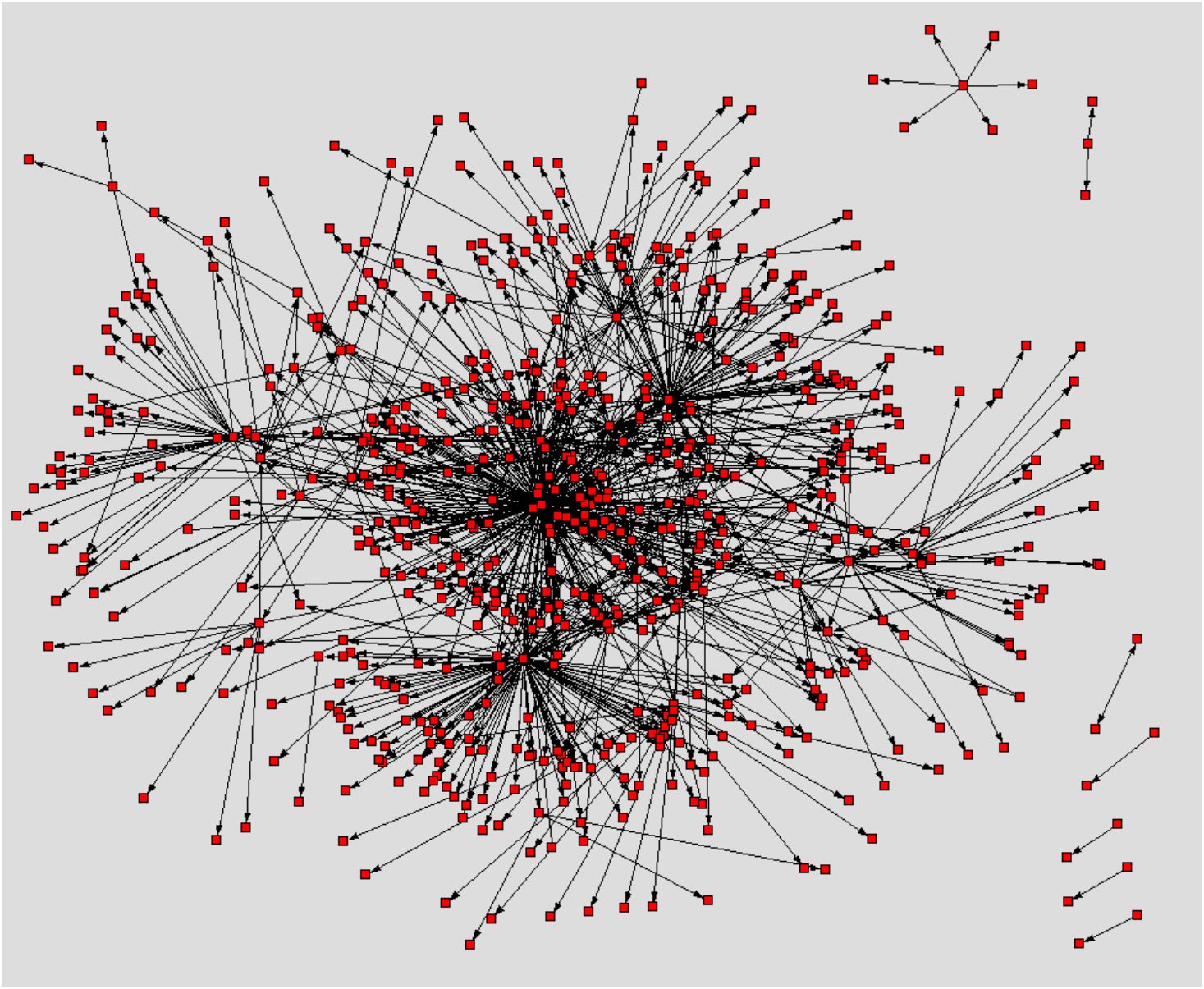
There are 1,947 individuals listed as authors or recipients.



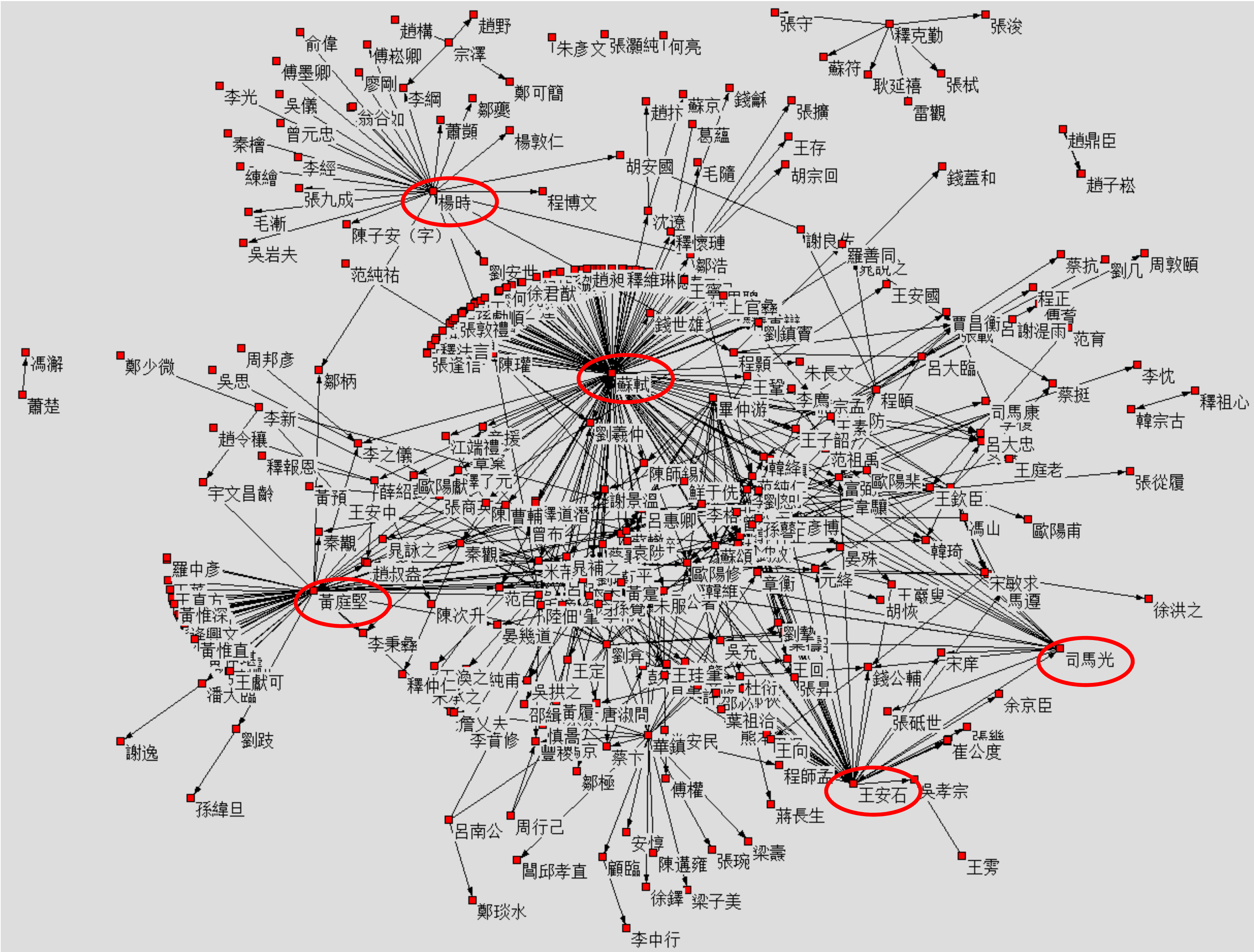
We also can look at NETWORKS of letter-writers: was everyone writing to everyone else or were some individuals at the center of the networks?



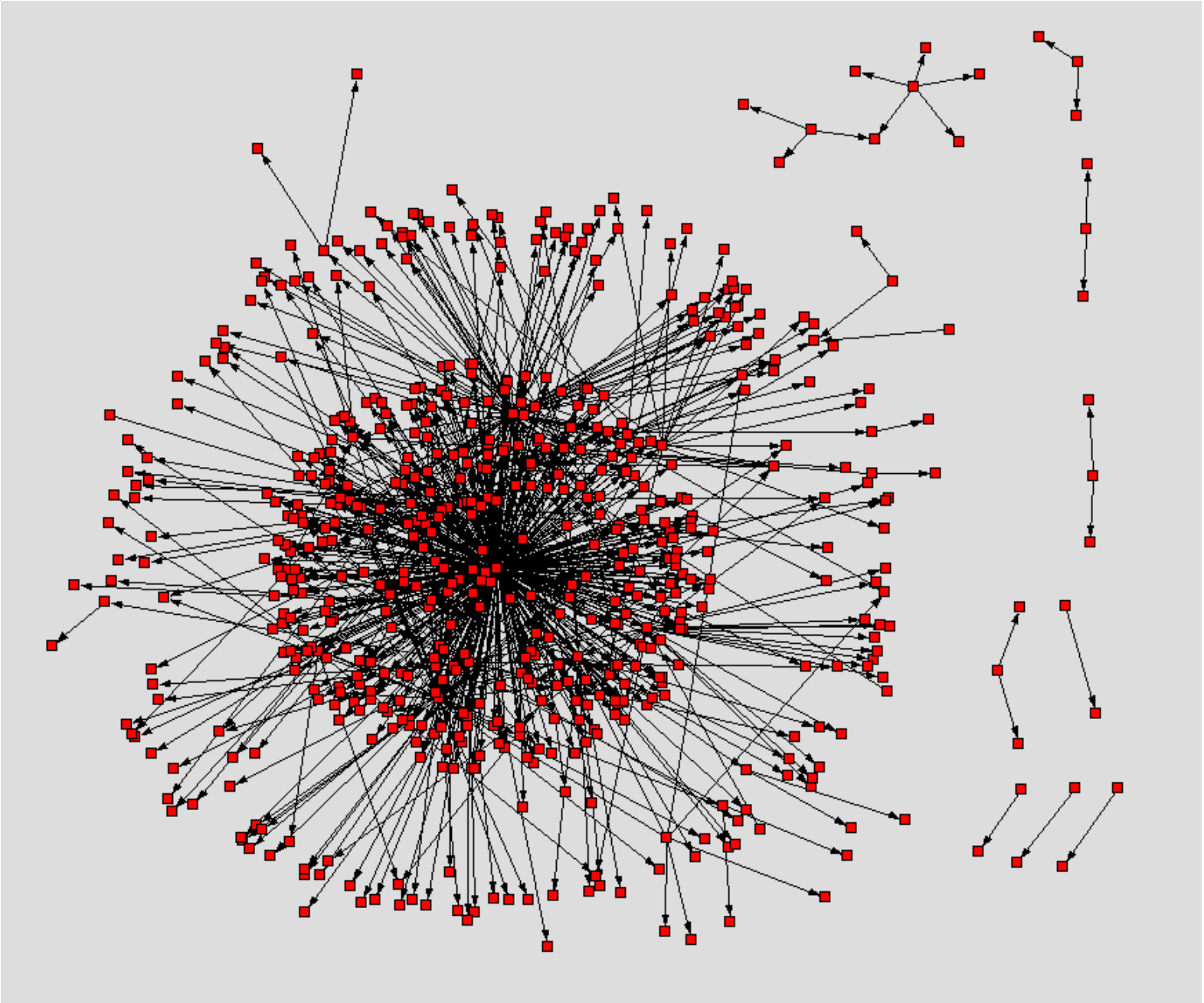
If we look at a fifty-year period, 1076 to 1125, determined by author index year:



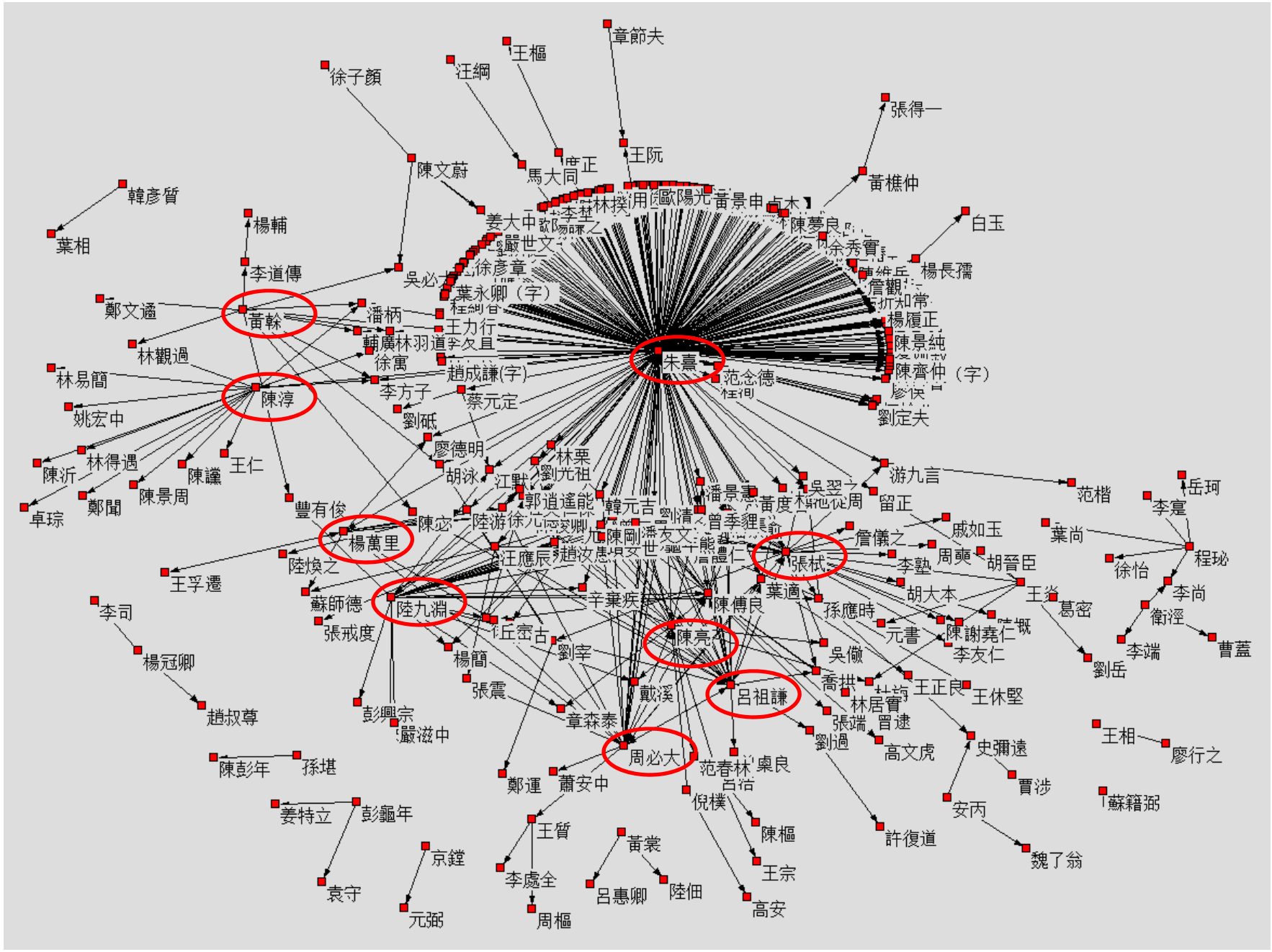
Looking in details, we get:



If we repeat this for 100 years later (1176-1225), we get the components:



We find the “usual suspects” at the center:



If we look at how many people were writing to **relatives**,
we get a slightly surprising result: