由史料中探勘社會網絡：以乾隆時期為例

Social Network Mining from Historical Documents– by Example during Qianlong's Reign

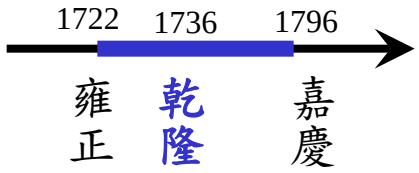
沈錳坤

Department of Computer Science, NCCU

研究概述

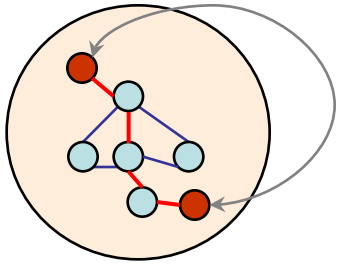
- 研究目的：由史料中探勘分析
  - 找出「權臣 (the chief counselors) 」
  - 判斷權力結構的改變
    - 權臣：沒落 ←→ 崛起

- 文本
  - 《清高宗實錄》：乾隆
    - 由盛轉衰 → 適合
    - 雍正13年 (1735) → 嘉慶4年(1799)，共65年
    - 官書：皇帝、中央政府



相關研究

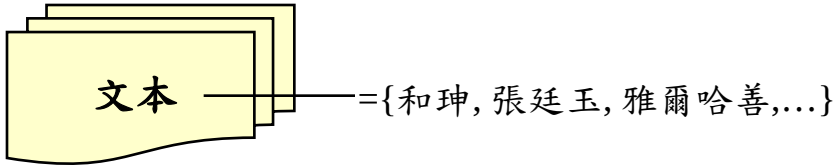
- 社會網絡分析 (Social Network Analysis)
  - Network Centrality [Freeman'79]
    - 量化指標 → 網絡節點重要性
    - Degree Centrality：網絡參與程度
    - Closeness Centrality：時間或成本─溝通整體網絡
    - Betweenness Centrality：橋樑─其它節點間連絡管道
  - Cohesive Subgroups [Wasserman'94]
    - 緊密相連的群體：密切關係
    - Based on Reachability and Diameter
      - *n*-cliques、*n*-clans 及 *n*-clubs
    - Based on Nodal Degree
      - *k*-plexes 及 *k*-cores



研究流程

Phase 1.

歷史人名識別

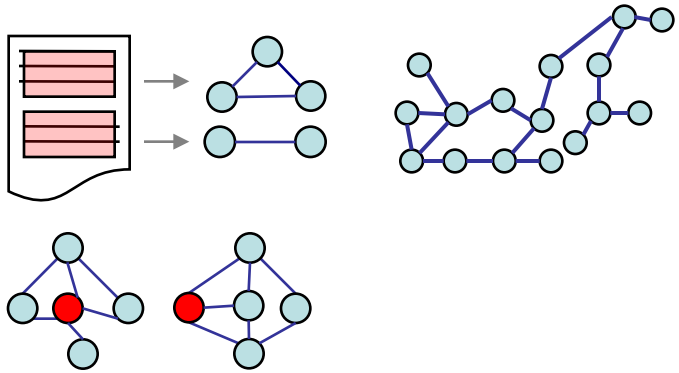


Phase 2.

探勘權臣

從文本建立人物間連結

建置當時的人脈網絡



在網絡中找出權臣

Phase 3.

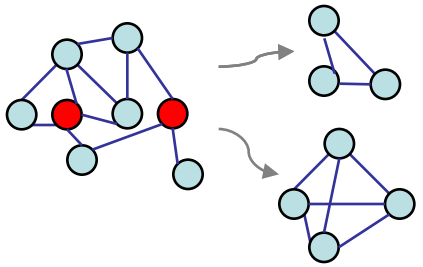
偵測權力結構的變化

方法一：

權臣個人的重要性變化

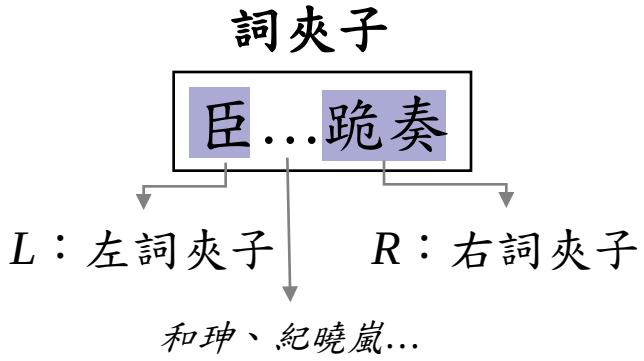
方法二：

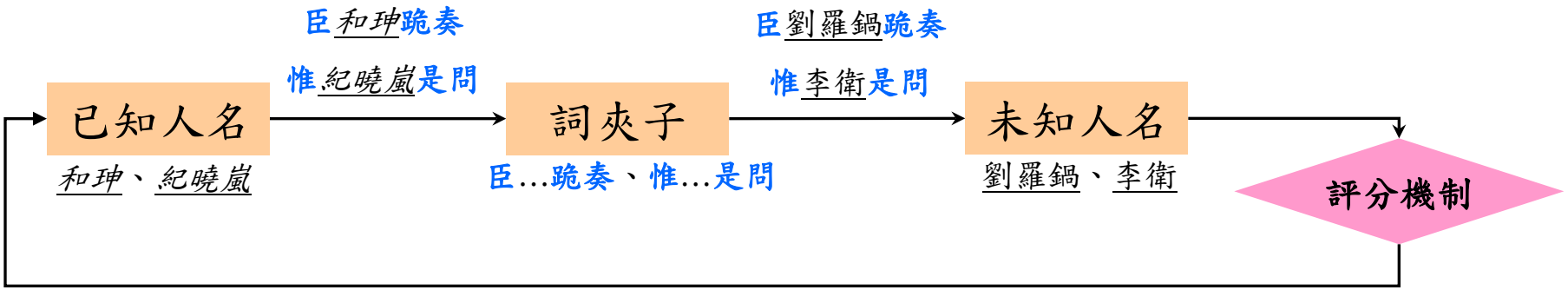
權臣的權力團體之消長



歷史人名識別

- 詞夾子演算法 [張尚斌’05]
  - 歷史文本時常有特殊的patterns
    - 「臣…跪奏」：臣和珅跪奏、臣紀曉嵐跪奏
  - 詞夾子 (Word-Clip)
    - 人名周遭的詞彙
    - 構成：***L*…*R***
  - 演算法核心





歷史人名識別 (續)

- 詞夾子演算法
  - 評分機制
    - **詞夾子分數** $\frac{R^2}{T}$
      - R：詞夾子夾中的樣本詞數
      - T：詞夾子夾中的總詞數
      - 平方：**好的詞夾子** → 夾中許多不同詞
        - » 詞夾子 A：夾中10個詞 (5個樣本詞)
        - » 詞夾子 B：夾中2 個詞 (1個樣本詞)

A $\frac{5}{10} \longrightarrow \frac{25}{10}$ 佳

B $\frac{1}{2}$ 劣

    - **人名候選詞分數** $Sc = \sum_{i=1}^{n} Swc_i$
      - Swc_i 為詞夾子 wc_i 的分數
      - $wc_1, wc_2, \ldots, wc_n$ 夾中候選詞 c
  - 為候選詞加分
    - 百家姓*I、官名*II

I

劉羅鍋
李衛

II

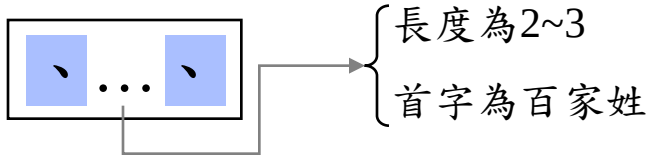
宰相劉羅鍋
李衛總督

歷史人名識別 (續)

- 修改詞夾子演算法
  - **左詞夾子必須是官名**：準確率↑、執行時間↓

    …*協辦大學士尚書阿桂*、*程景伊*、*署協辦大學士尚書英廉*、*尚書豐昇額*、*袁守侗*、*福隆安*、*綽克托*、*奎林*、*侍郎福康安*、…〈清高宗實錄1028卷〉

  - **左右詞夾子皆為頓號**：召回率↑

    …*所有列在一等之進士舉人邱桂山*、*祝盆*、*洪榜*、*戴衢亨*、*關槐*、*俱著以內閣中書補用*。…〈清高宗實錄1007卷〉

- 過濾錯誤候選詞
  - 地名*I、官名*II、普通詞庫*III
  - 長期出現在文本中*IV
    - 60年

I	II
江南 蘇州	左庶子 武備院卿
III	**IV**
情形 前往	奏稱 議覆

探勘權臣

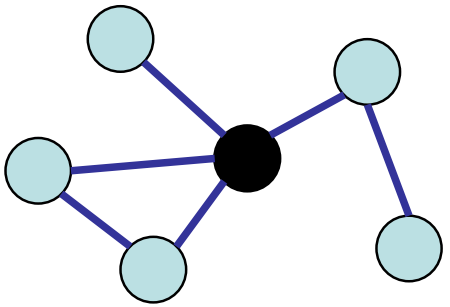
- 為歷史人物間建立連結
  - 網絡的表達
    - Node：人物
    - Link：人物間存在關係 (連結)
  - 文本符號「○」

    …1*○辛未。上詣皇太后宮問安。*2*○遣官祭關帝廟。*3*○調原任浙江巡撫黃叔琳。直隸按察使浦文焯。*…〈清高宗實錄1012卷〉

    - 時序排列：收錄諭旨、奏疏 → 事件
  - 建立：位於相同「○」

- 人脈網絡的建置
  - 方式：unweighted 及 weighted
  - 單位：年

探勘權臣 (續)



p_k

- Network Centrality
  - **Degree Centrality $C_D(p_k)$**：節點分支度
    - 參與很多政務，與多人產生連結

$$C_D(p_k) = \sum_{i=1}^{n} a(p_i, p_k)$$

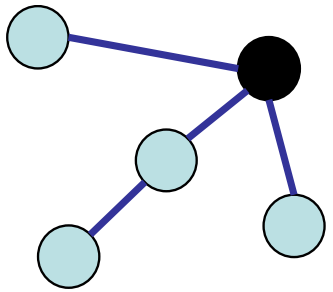
- p_i：網絡上的任一點
- $a(p_i, p_k)$：點 p_i 與點 p_k 相鄰 → 是：值；否：0

  - **Closeness Centrality $C_C(p_k)$**：拜訪網絡的最短路徑
    - 掌握網絡有良好能力

$$C_C(p_k) = \sum_{i=1}^{n} d(p_i, p_k)$$

- $d(p_i, p_k)$：點 p_i 與點 p_k 的最短路徑長度

    - 使用 Floyd-Warshall Algorithm

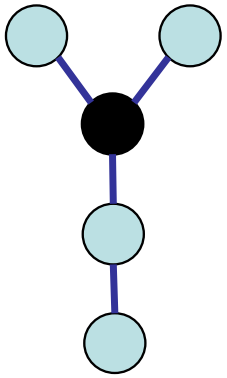


探勘權臣 (續)

p_k • Network Centrality

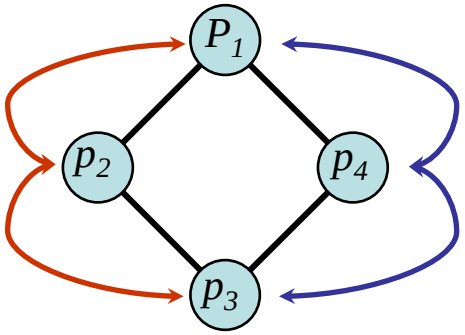
– **Betweenness Centrality $C_B(p_k)$**：資訊流通時經過的頻率

- 溝通派系 → 身處其它人物彼此溝通的特殊角色
- 最短路徑：$p_1 \longleftrightarrow p_3$
  - p_2、p_4 **0.5**



$$C_B(p_k) = \sum_{i<j} b_{ij}(p_k)$$

遞減排序



- Floyd-Warshall Algorithm

偵測權力結構的變化

- 根據兩個不同面向
  (1)權臣重要性之變化
  (2)權力核心團體之變化

(1) 基於權臣個人的重要性變化

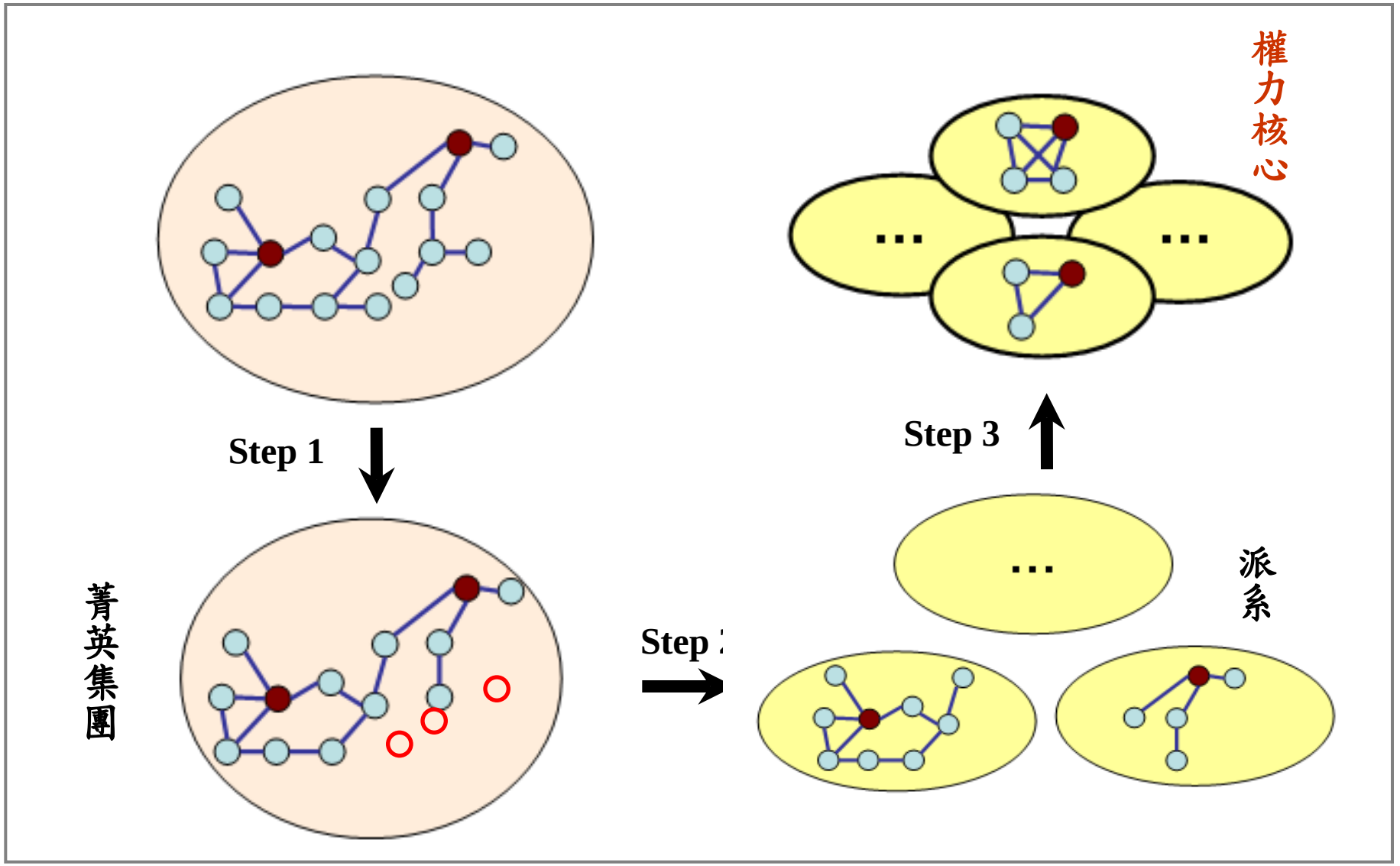
- 權臣重要性：Centrality
- 每年所有權臣之重要性： vector space model
- 兩年間的權力結構變化：Difference of vectors

$$\left\{ Diff(Y_{N-1}, Y_N) = \sum_{i=1}^{m} \left| C_{N-1}(i) - C_N(i) \right| \right.$$

- $C_{N-1}(i)$：權臣 i 第 $N-1$ 年的 Centrality 值
- $C_N(i)$：權臣 i 第 N 年的 Centrality 值

偵測權力結構的變化 (續)

(2)權力核心團體之變化

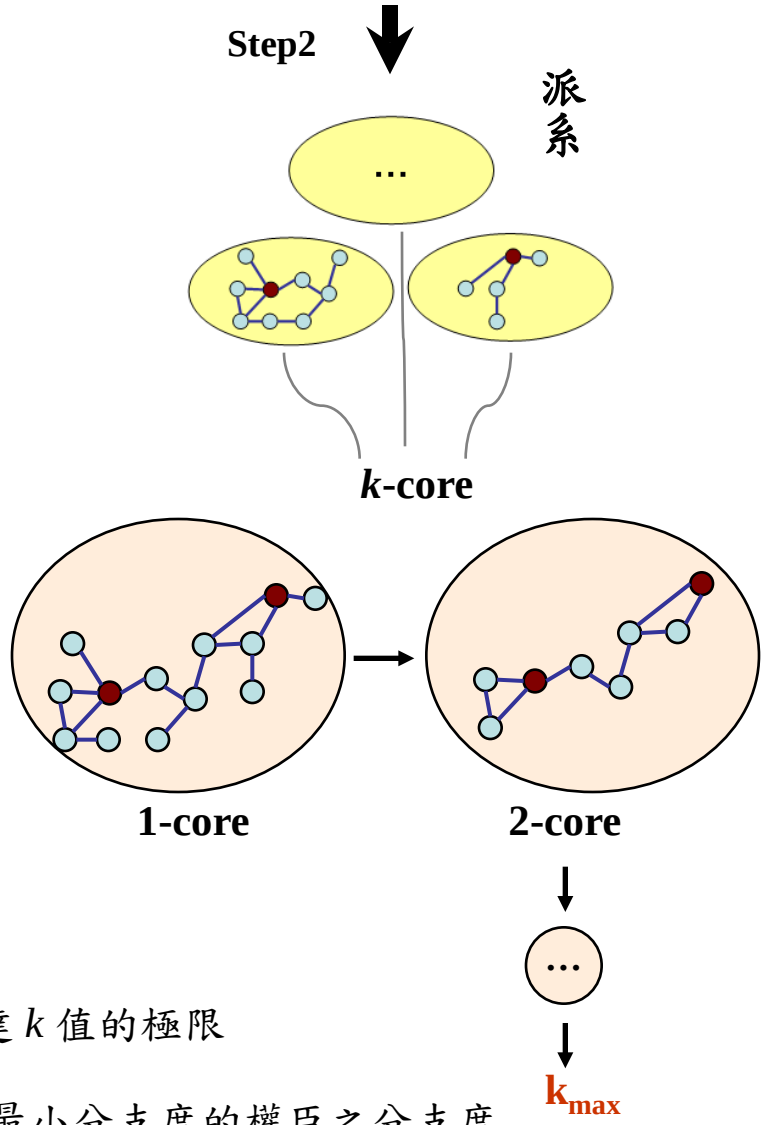


偵測權力結構的變化 (續)

- 基於權力團體的消長
  - 尋找的流程
    - **Step 2：派系** — **用 *k*-cores 分解**
    - Subgroups 內與多少成員相鄰
    - 若 N_s 為 *k*-core

$$d_s(i) \geq k \quad \text{for all} \quad n_i \in N_s$$

    - 從 *k*=2
      - *k*=1 → connected component
      - ***k*↑ → Subgroups 更緊密**
    - 停止條件
      - 目前k-cores 的 k 值

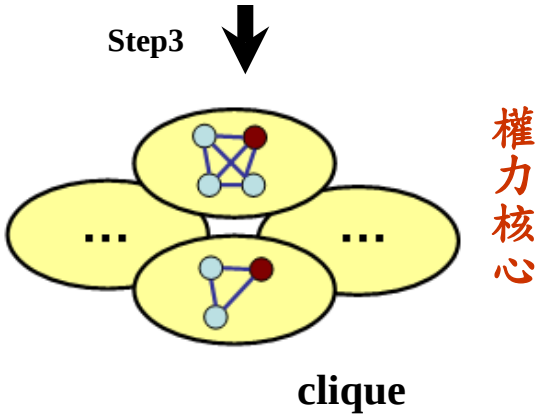


$$k_{max} = dg(Chief_q) \quad \rightarrow \text{達 } k \text{ 值的極限}$$

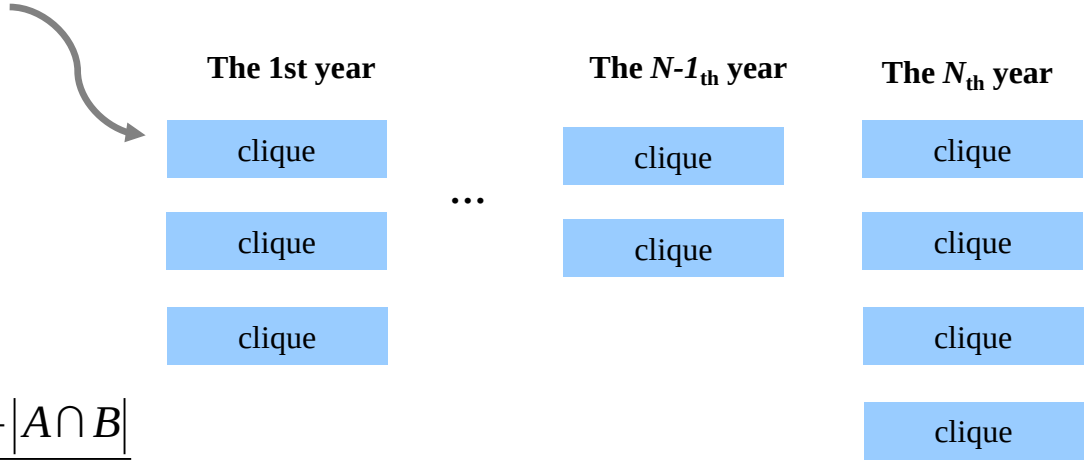
- $dg(Chief_q)$：目前 *k*-cores 內，有最小分支度的權臣之分支度

偵測權力結構的變化 (續)

- 基於權力團體的消長
  - 尋找的流程
    - **Step 3：權力核心─ 尋找clique**
    - 權力團體─ 人物間的關係最 cohesive
    - Maximal cliques



  - 消長：團體差異
    - Cliques 間差異
    - Jaccard distance



$$J_\delta(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

- $|A \cup B|$：為集合A與集合B聯集的個數
- $|A \cap B|$：為集合A與集合B交集的個數

偵測權力結構的變化 (續)

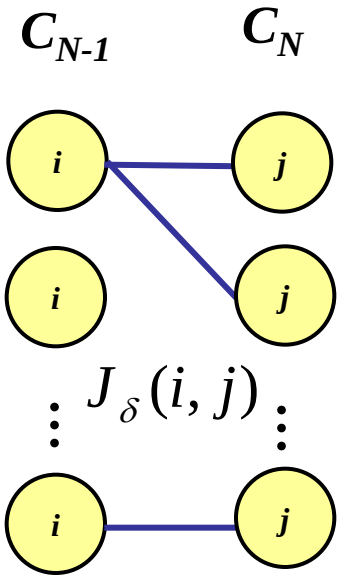
- 基於權力核心的消長
  - 兩年間權力結構變化量

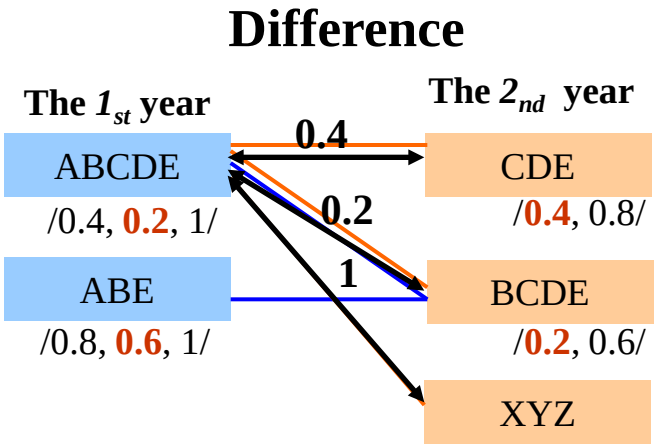
$$Diff(C_{N-1}, C_N) = \left(\sum_{i, i \in C_{N-1}} \min_{j, j \in C_N} J_\delta(i, j) + \sum_{j, j \in C_N} \min_{i, i \in C_{N-1}} J_\delta(j, i) \right) \Big/ \left(|C_{N-1}| + |C_N| \right)$$

標準化

- i為第N-1年權力核心內的某個clique
- j為第N年權力核心內的某個clique

> **差異門檻值** → 判斷有所變化!





$Diff(C_{1st}, C_{2nd})$
= (0.2+0.6+0.4+0.2+1) /5 = 0.48

實驗

- 評估方法：Precision、Recall 及 F-score

$$P = \frac{\text{Number of relevant items retrieved}}{\text{Total number of items retrieved}}$$

$$R = \frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items in collection}}$$

$$F = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R}$$

所有實驗皆 **β=1**

- 歷史人名識別
  - 全1500卷 → 隨機標注15卷 (1%)
  - 實驗結果(1)：限定左詞夾子為官名

	P	***R***	***F***	***Num***
NER(0)	11.36%	41.71%	**17.86%**	44,196
NER(1)	51.81%	39.41%	**44.76%**	6,779

without

with

實驗 :人名識別

- 歷史
  - 實驗結果(2)：利用詞庫過濾候選詞

	P	*R*	*F*	*Num*	
NER(1)	51.81%	39.41%	**44.76%**	6,779	**without**
NER(2)	84.63%	37.25%	**51.73%**	5,872	**with**

  - 實驗結果(3)：利用頓號補召

	P	*R*	*F*	*Num*	
NER(2)	84.63%	37.25%	**51.73%**	5,872	**without**
NER(3)	77.54%	44.91%	**56.87%**	10,157	**with**

  - 實驗結果(4)：剔除長期出現在文本的詞彙 (60年)

	P	*R*	*F*	*Num*	
NER(3)	77.54%	44.91%	**56.87%**	10,157	**without**
NER(4)	78.08%	44.76%	**56.90%**	10,141	**with**

實驗:探勘權臣

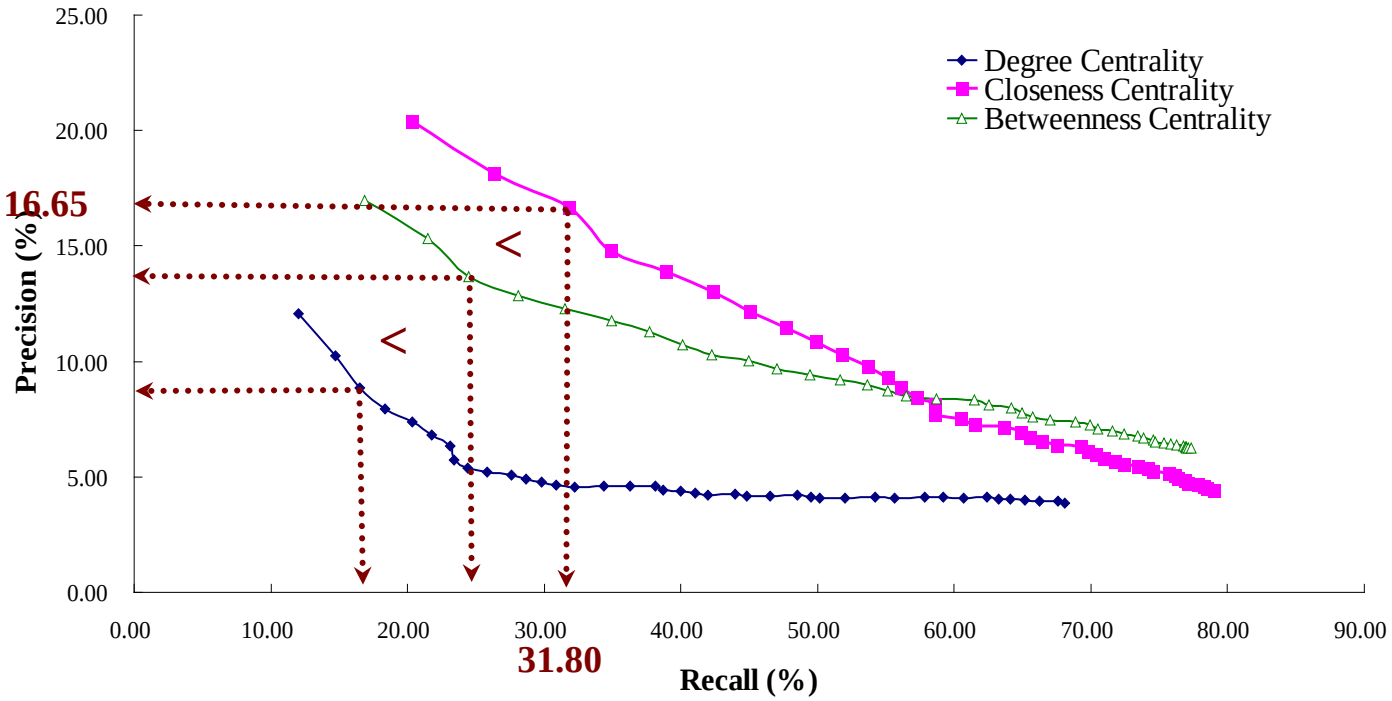
- 探勘權臣
  - Ground truth: **軍機大臣**及**內閣大學士**

*...自清世宗雍正設立軍機處以來，內閣權傾，**軍機處**漸重，已成為清代中央政府中具有重要威權的機構...*[古鴻廷’05]

*...清初沿襲明制，雍正以後，始設軍機處，至是，內閣無實權，然**大學士**亦常入軍機，固仍不失為宰輔也。...*[蕭一山’62]

實驗:探勘權臣(續)

– 實驗結果：網絡中心性間的比較 (unweighted)



掌握

斡旋
→ 君權、嚴密

活絡
→ 中高階

實驗:探勘權臣(續)

- Closeness Centrality：*n*+10
  - Precision = 16.65%　Recall = 31.80%　F-score = 21.86%
    - 低估準確率

雍正13年部份結果

不在標準答案集

傅鼐：副都統銜(正二品)

甘汝來：禮部右侍郎(從二品), 副總裁, 兵部尚書(從一品)

三泰：協辦內閣大學士(從一品)

任蘭枝：吏部左侍郎(從二品), 世宗憲皇帝實錄總裁官, 禮部尚書(從一品) 等

福敏：協辦大學士(從一品), 太子太保(正一品), 翰林院掌院學士(從二品) 等

徐元夢：內閣學士(從二品), 刑部右侍郎(從二品), 禮部右侍郎(從二品) 等

張廷璐：工部右侍郎(從二品), 世宗憲皇帝實錄副總裁官 等

高

正一品

從一品

正二品

…

從九品

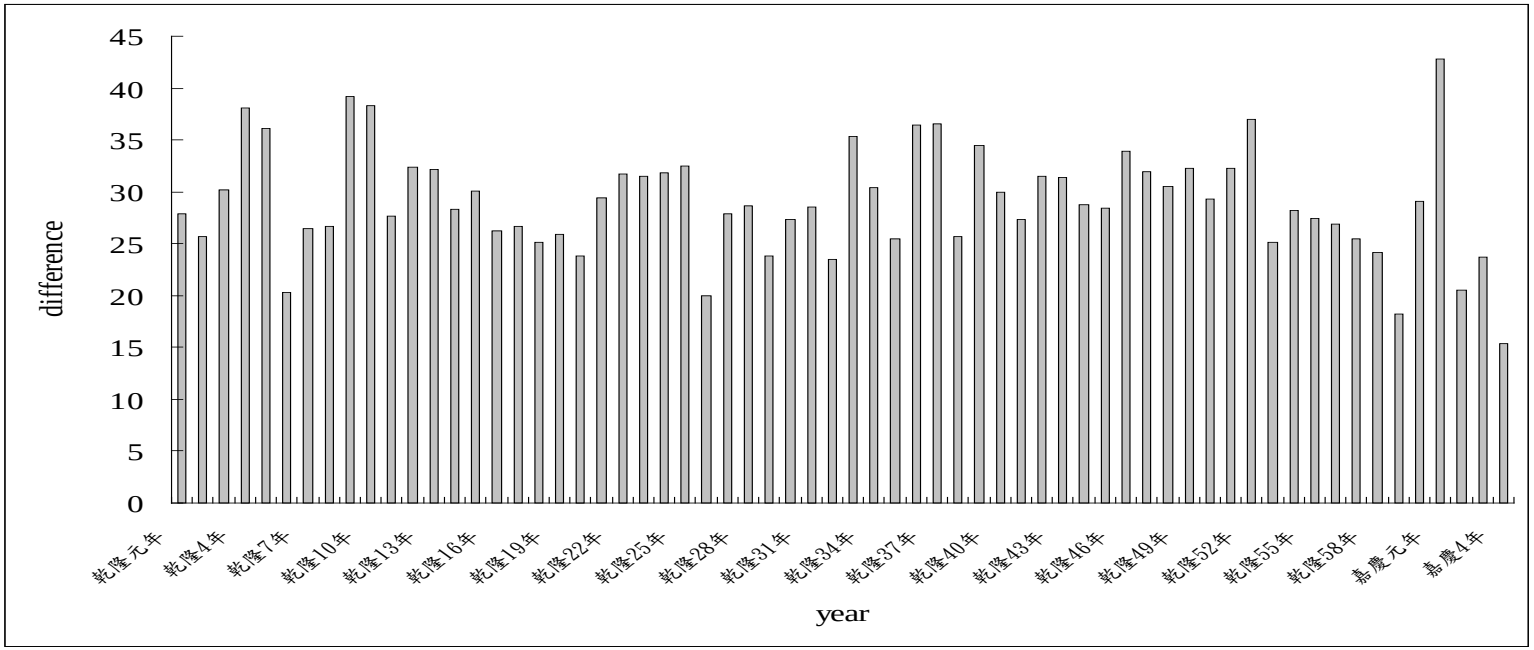
低

實驗:偵測權力結構變化

- Ground truth: 軍機領班大臣解職時間

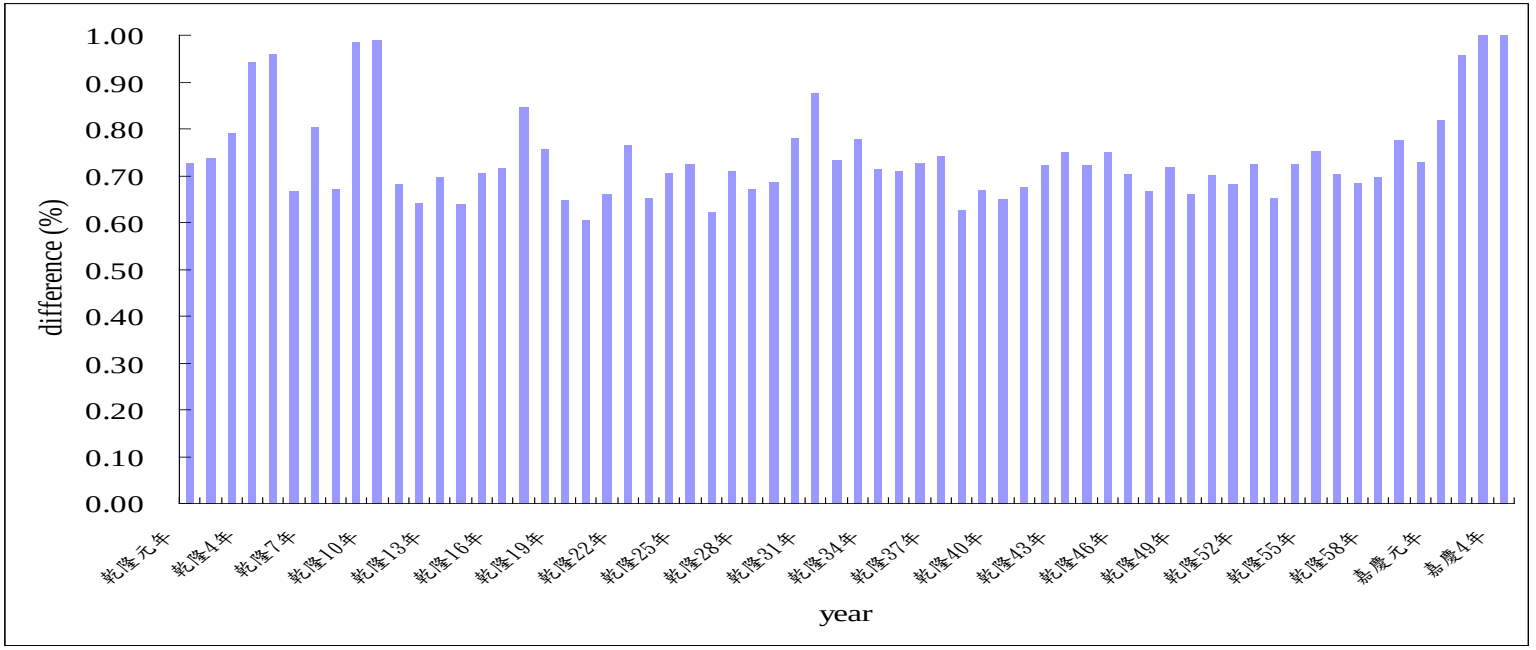
*…軍機處創立後，內閣權輕，時人遂改以**軍機領班大臣**為**相權**之代表…*[蔡秉叡’07]

– 實驗結果(7)：兩年間權力結構差異– 基於權臣個人的重要性變化



實驗 (續)

- 偵測權力結構變化
  - 實驗結果(8)：兩年間權力結構差異—基於權力團體的消長



  - 實驗結果(9)：權力結構偵測的效果

	P	***R***	***F***
基於權臣個人的重要性變化	41.67%	55.56%	**47.62%**
基於權力團體的消長	57.14%	44.44%	**50.00%**

實驗:偵測權力結構變化(續)

- 基於權力團體的消長
  - Precision = 57.14%　Recall = 44.44%　F-score = 50.00%
- 亦低估準確率
- 標準答案：軍機領班大臣的解職
  - **內閣大學士** [古鴻廷’05]

解職— 死亡、貶職

...內閣大學士為清代官僚體系中的最高職位，陞及此人臣之極的正一品官職，自非易事。內閣大學士往往為軍機處成員，而軍機大臣之職位僅為一項兼職...

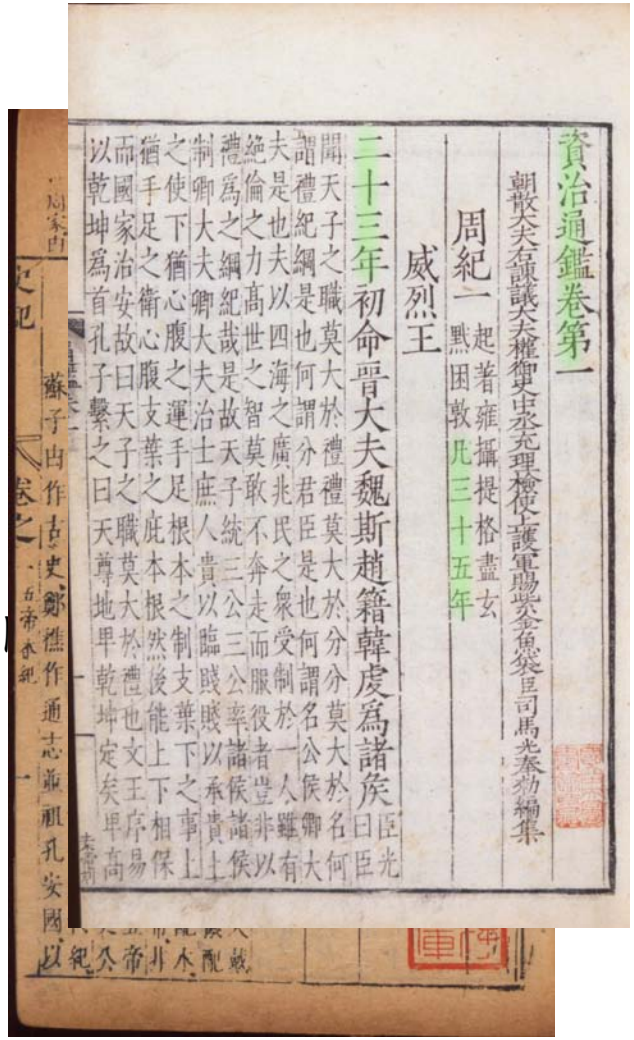
年代	
乾隆5年	
乾隆11年	**正確答案**
乾隆31年	→ **莊有恭 解職 (內閣大學士)**
乾隆36年	**正確答案**
乾隆45年	**正確答案**
乾隆58年	
嘉慶3年	**正確答案**

未來研究

- 歷史資訊學？協助歷史研究之工具
- Recall 重於Precision
- Ongoing research
  - 根據「一人得道雞犬升天」探勘派系community mining
  - 編年體 vs. 紀傳體 vs. 編年事件別史之轉換
  - 探勘史學矛盾
  - 計量歷史學

紀傳體編年體轉換

- 紀傳體轉換編年體
  - 紀傳體
    - 以人物為中心
    - 《史記》等二十四史
  - 編年體
    - 按時間先後次序
    - 《左傳》、《資治通鑑》
  - 編年事件別史？
    - 某一領域的歷史
    - 《晚清帝國主義侵入史》、《大清與外族
    - 加入其他「二十四史」後
      - 針對明清兩代
        - » 以《明史》與《清史稿》做材料
        - » 《中國前近代社會與文化史》
      - 以《史記》、《三國志》、《五代史》
        - » 《中國的分裂時代》
        - » 分析比較出中國分裂時代的特徵 等



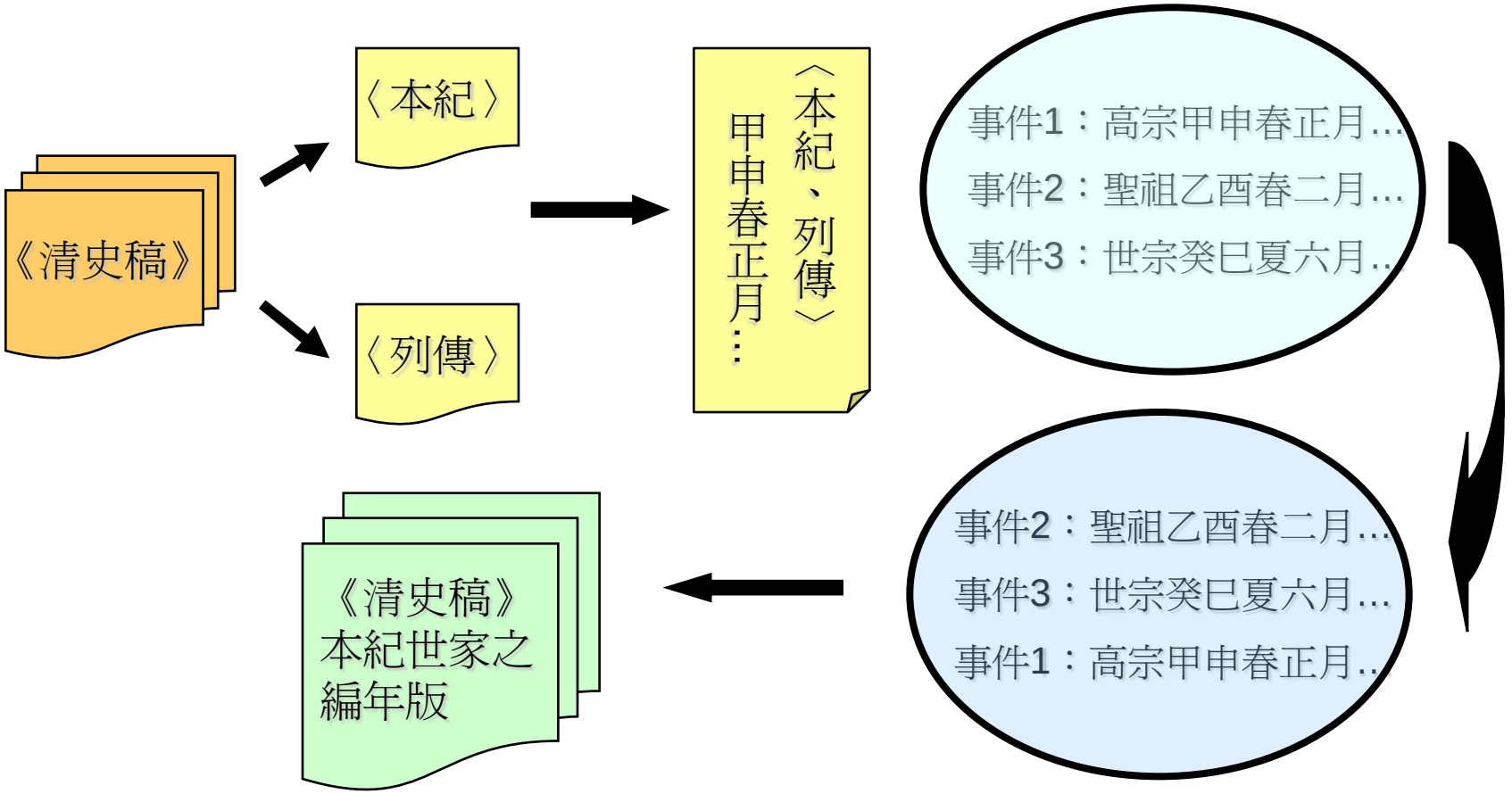
紀傳體編年體轉換(續)

- 編年體轉換紀傳體？
  - [反面]
    - 各代已存有紀傳形式的正史
    - 實用性？
  - [正面]
    - 尋找矛盾點
    - 史學家早已知道？
    - 過去史家撰史的材料
      - 正史之外？

紀傳體編年體轉換(續)

- 從事件文句中挖掘隱含訊息？
  - 事件關聯法則
    - 「天災」與「罪己詔」
    - 主軸研究後下一步！
    - 其他可能應用？

紀傳體編年體轉換(續)

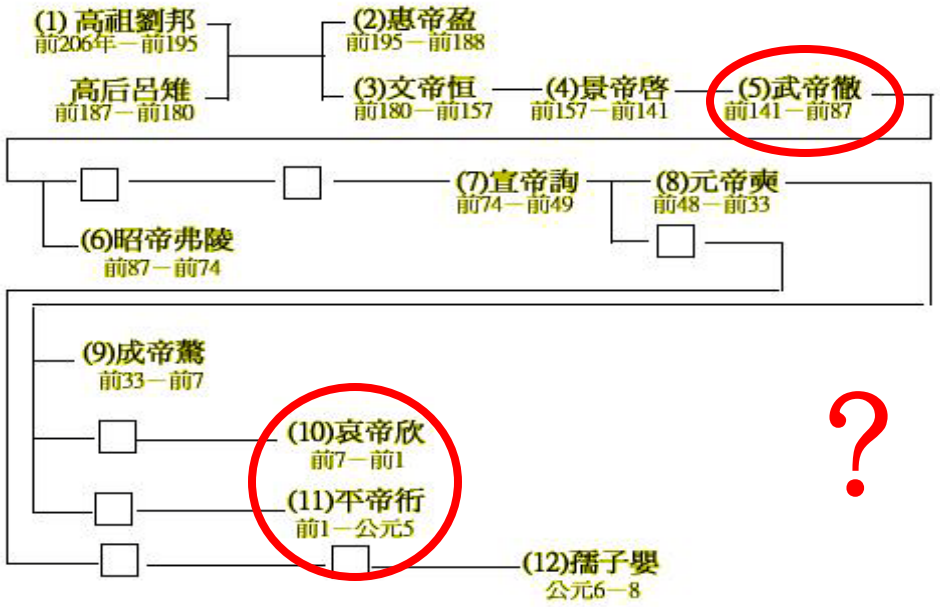

《清史稿》
〈本紀〉
〈列傳〉
〈本紀、列傳〉
甲申春正月…
事件1：高宗甲申春正月…
事件2：聖祖乙酉春二月…
事件3：世宗癸巳夏六月…
事件2：聖祖乙酉春二月…
事件3：世宗癸巳夏六月…
事件1：高宗甲申春正月…
《清史稿》
本紀世家之
編年版

史學研究：比較方法

- 矛盾(《史記》〈司馬相如傳〉贊)
  - 「相如雖多虛辭濫說，...揚雄以為靡麗之賦，勸百風一，...。」

司馬遷是武帝時代的人

→ 揚雄是哀帝、平帝、王莽時代的人！



史學研究：比較方法 (續)

- 除了找尋矛盾...
  - 避諱於本紀，散見於列傳
  - 「《三國志》雖多迴護，而其翦裁斟酌處，亦自有下筆不苟者。...郭后李陰貴人，竝愛幸，甄失志，出怨言，帝怒，遂賜死。是雖諱之於紀，猶載之於傳也。」
    (趙翼《廿二史劄記》卷六)

→ 發現歷史的真面目！

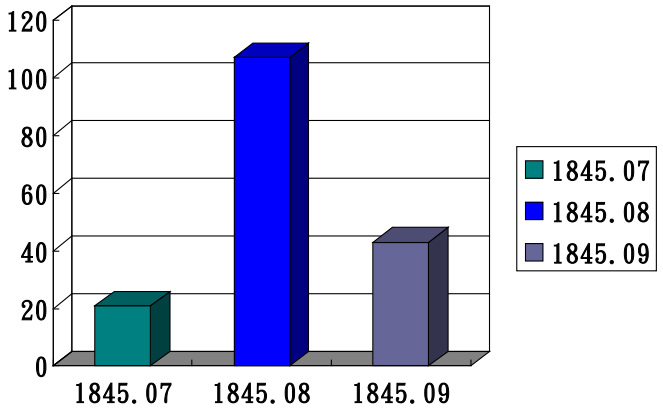
史學研究：比較方法 (續)

- 轉手記載與原書比較
  - 一手史料 v.s. 二手史料
  - 抄錄、刪節、潤飾、斷章取義、竄改
  - 《資治通鑑》
  - 《廿二史劄記》
    - 「吳〈孫輔傳〉，其子松為射聲校尉都鄉侯，黃龍三年卒。蜀相諸葛亮與兄瑾書曰...。」(〈三國志誤處〉)

→ 《三國志》〈孫翊傳〉

史學研究：計量分析法

- 用形容詞、動詞出現探討
  - 歷史人性
  - 發掘時代
- 海爾 (William Bayard Hale, 1856-1924)
  - 美國總統─威爾遜
- 波德(David P. Boder)
  - 形容詞與動詞商數(A.V.G)
  - 美國哲學家愛默森 (R.W.Emerson, 1803-1882)



史學研究：計量分析法(續)

- 計量歷史學？
  - 「名詞消長」、「新事物普及」
  - 一件事物何時萌芽、盛行、消失？
    - 「胡風與胡化」
    - 「基督教在中國歷史上的消長」
  - 沒有資訊科技的幫助
    - 只能透過傳統研究法
      - 上泉碧落下黃泉！