

大數據分析與數位人文 對圖書資料之意涵



劉吉軒

創新與創造力研究中心
資訊科學系

2017臺灣博碩士論文雲端書庫全國聯盟年會

大綱

- 基本脈絡概念
- 應用案例
- 功能意涵
- 圖書館意義總結

大數據的核心定位

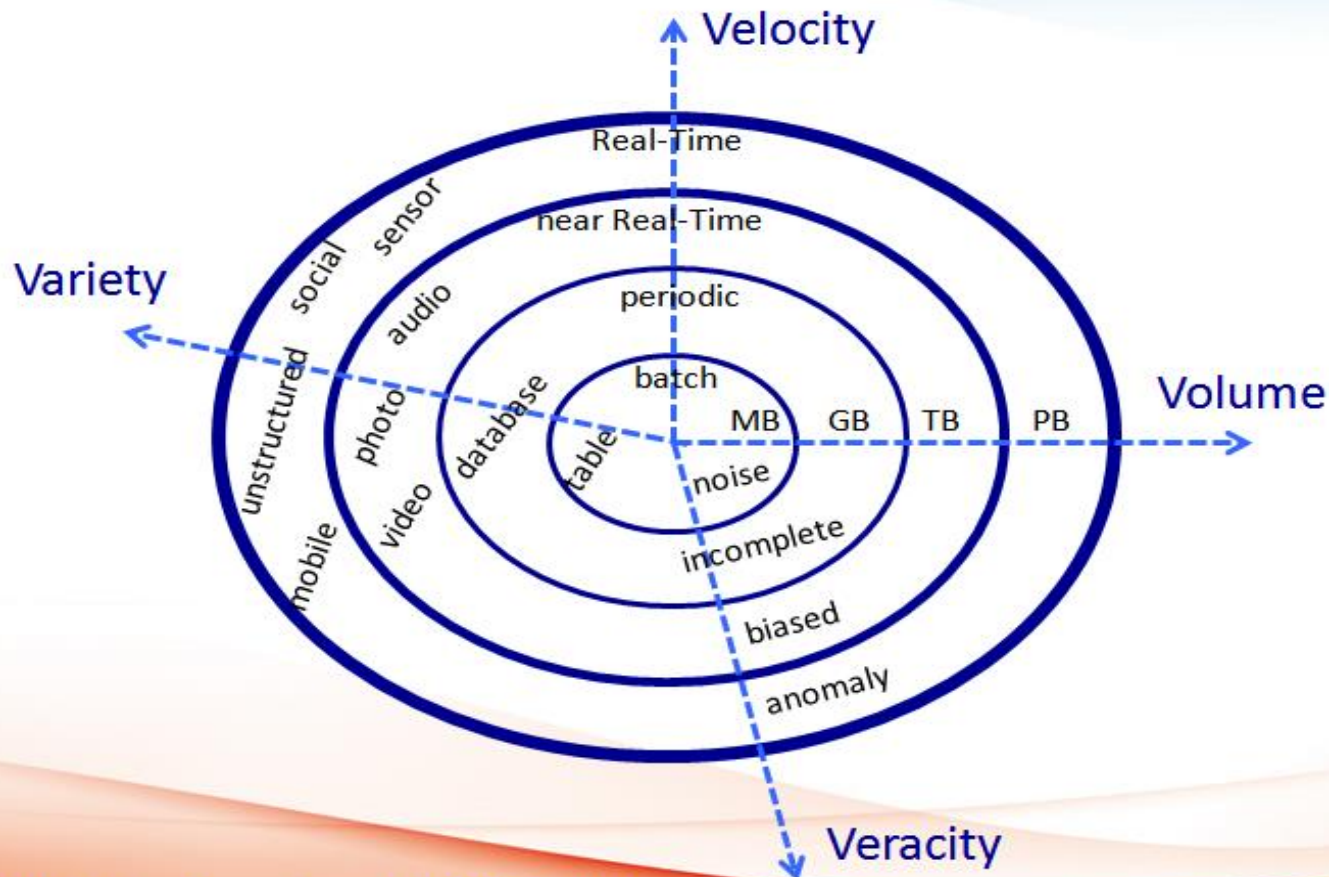
- 發覺顯著差異與關連
- 解讀規律型態
- 瞭解過去
- 預測未來



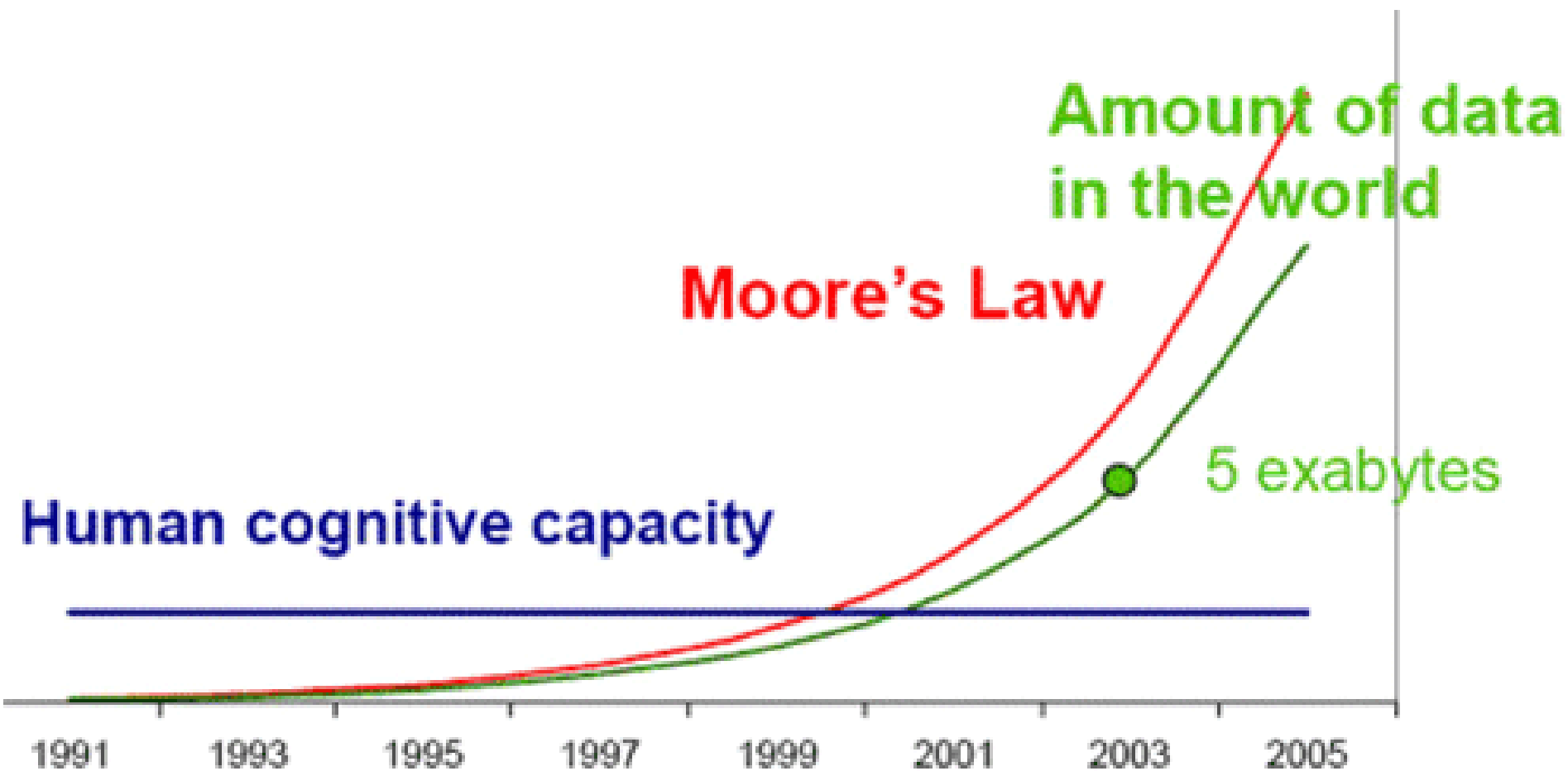
★ 關鍵：掌握資料、善用資料!

大數據的光譜

Big data Expands on 4 fronts

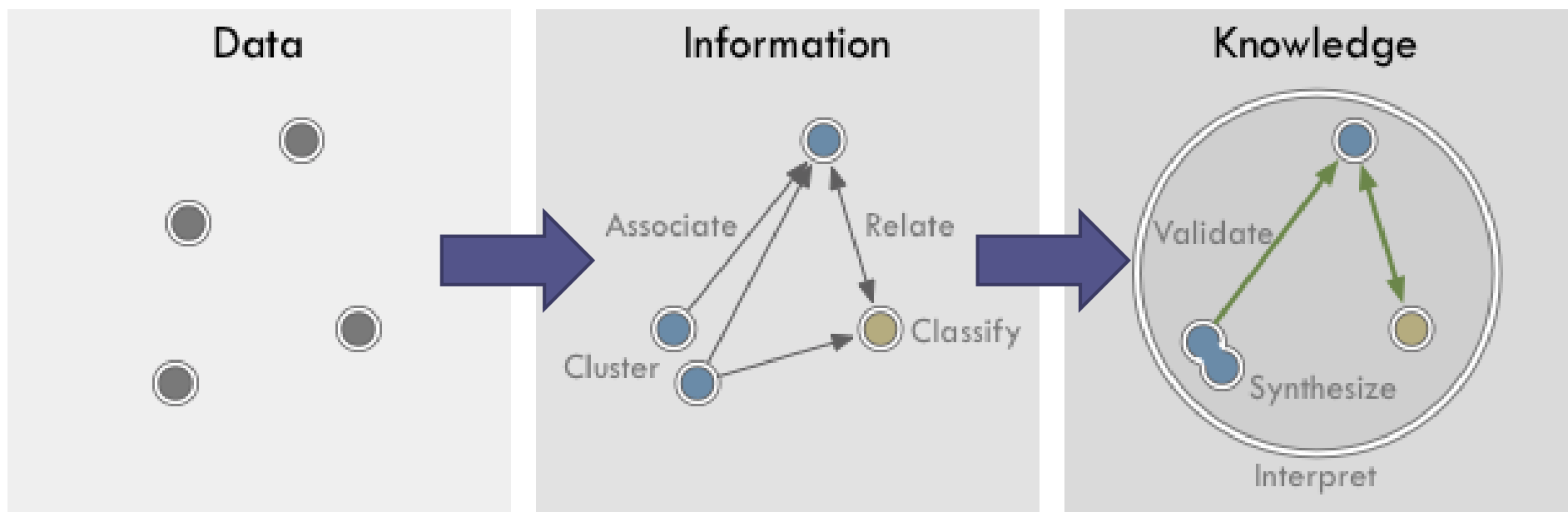


人腦的大數據 vs. 機器的大數據



大數據的計算觀點脈絡

- 資訊科學 (Computer Science/Information Science) 的長期發展主軸 – **建立系統性連結**



Decision **Action**

大數據分析方法

- 資料探勘 (Data Mining)

- 對大規模資料進行自動或半自動的分析，以提取過去未知的有價值的潛在資訊
- 能達到以上目的的多種技術的集合; 從統計, 演算法到機器學習, 類神經網路, 深度學習, 強化學習...等, 都可能被運用

- 非監督方法 (unsupervised)

- 在不具備認知的情況下, 對資料進行探索 (discovery with no prior knowledge)

- 監督方法 (supervised)

- 對部分資料具備已知答案, 嘗試讓電腦自己找出資料特徵, 能建立和人類一樣的判斷能力

文字是重要資料!

- 文字是人類語言的書寫模式
- 文字是表達與記錄的最重要媒介
- 文字涵蓋了任何面向的資訊與知識
 - 從古至今
 - 從歷史, 文學, 到商業, 政治, 社會趨勢等

從大數據的角度看待文字
(80% 的資訊量)

文本探勘 (text mining)

- Textual patterns and relations
- 幾種典型功能
 - 總結 (summarization), 從keywords (關鍵詞)到文件摘要
 - 抽取觀念 (concept extraction), 如概念連結與類比關係
 - 計算文字特徵, 建立預測分析模型, 如主題分類, 作者辨識

自然語言處理 (natural language processing)

- 結合語言學領域知識, 數理統計, 邏輯演算, 認知科學等, 嘗試掌握人類語言的理解與應用能力
- 建立概念類比關係, 如 男人:女人 = 國王:???
- Japan: Sushi = German: ??? = America: ???
- 從部分文字輸入, 預測接下來可能出現的文字
- 建立語意概念網絡, 回答問題, 與人對話

網絡模型捕捉了許多複雜現象...



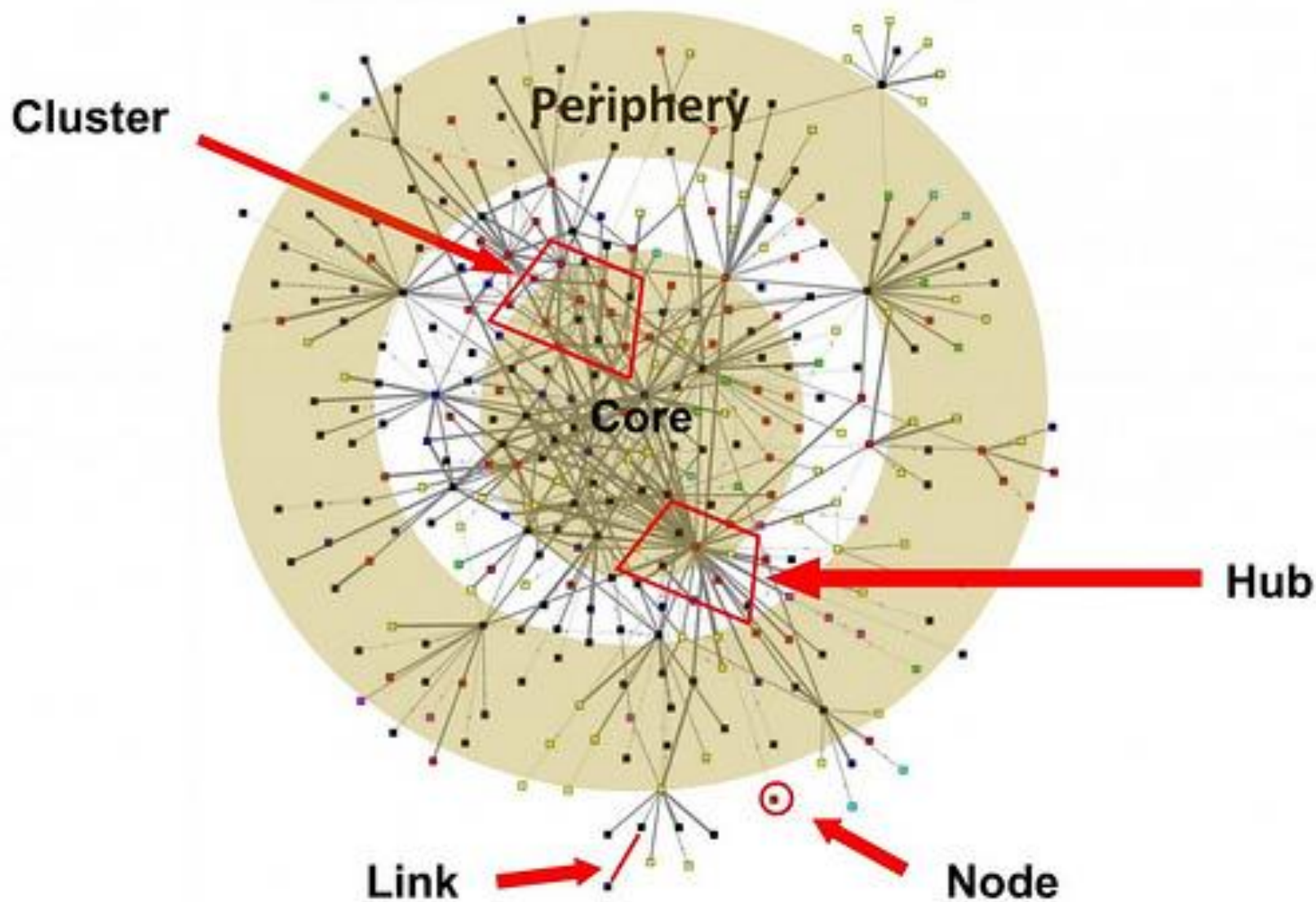
... 到處都是相互連結、相互影響的系統



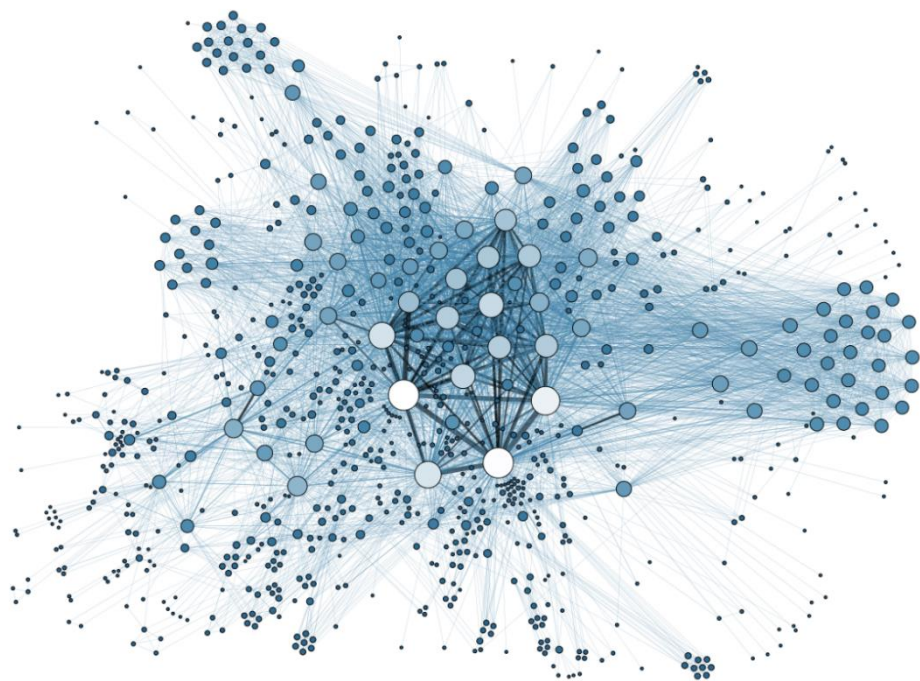
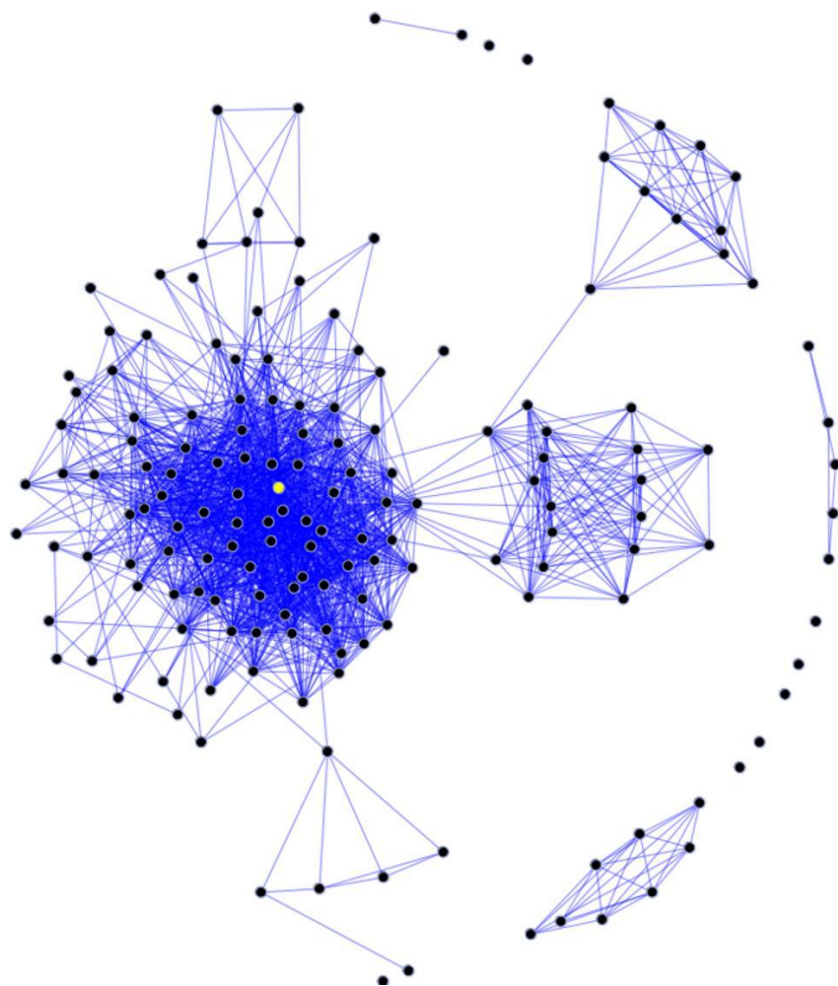
連結系統中的重要資訊

- 整體結構樣貌與差異
- 特殊個體
 - 個體的位置與角色
- 個體之間的路徑
- 子群組
 - 集團, 區塊, 群組, 派系, ...

找出結構內部特徵/差異

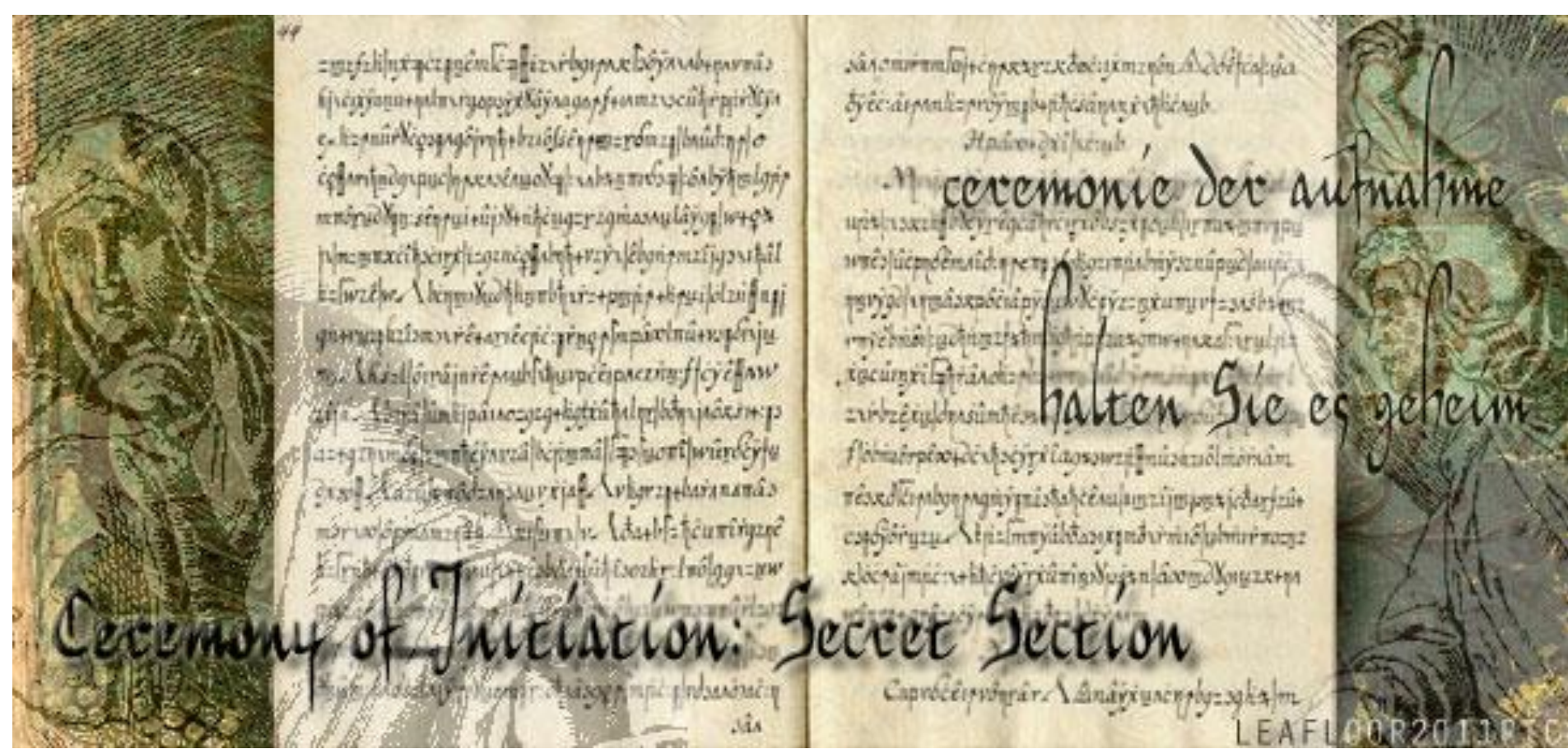


整體結構樣貌特徵與差異



An Example of Digital Humanities

How Revolutionary Tools Cracked a 1700s Code
(Oct. 24, 2011, New York Times)



解譯古代密碼書

- 一位歷史學家從解譯過的古書內容中, 發現討論天賦人權(natural rights of man)的政治言論, 比現有已知的文獻更早
- 對歷史學的重要貢獻
 - 政治思想(political ideas)擴散的過程
 - 十八世紀的歐洲, 秘密會社盛行, 對法國革命與美國革命都有重要影響

數位人文

- 數位科技與人文議題的交會與融合
 - Humanities computing – 1960年代啟動
 - 從數位典藏到數位人文
 - 各項資源條件更成熟後, 更多元、更大能量的開展
- 資料計算提供一個跨域框架的本體，數位人文是在此基礎框架上，對人文資料與人文議題的不同專業領域，如文學、歷史、語言、政治、經濟、傳播等，展開垂直與水平的各種跨域組合與交融

文本中的語言特徵

- 寫作風格分析與作者辨識
 - Shakespeare attribution studies
 - Harry Potter by J. K. Rowling vs. The Cuckoo's Calling by Robert Galbraith
 - 紅樓夢 前八十回 vs. 後四十回

經濟部商業司商工登記

- 台灣公司登記、商業登記、工廠登記等商工行政資料
 - 包括公司統一編號、公司名稱、資本額、董事會成員、各成員出資額或持股比率等
 - 定期公布, 已達三百萬筆以上資料
- 部分廠商開發商工行政資料整合查詢的服務
- 民間組織台灣零時政府（g0v）也利用視覺化技術，開發了互動式的台灣公司網路圖，可提供線上查詢探索，檢視特定公司與其他公司的依存關係

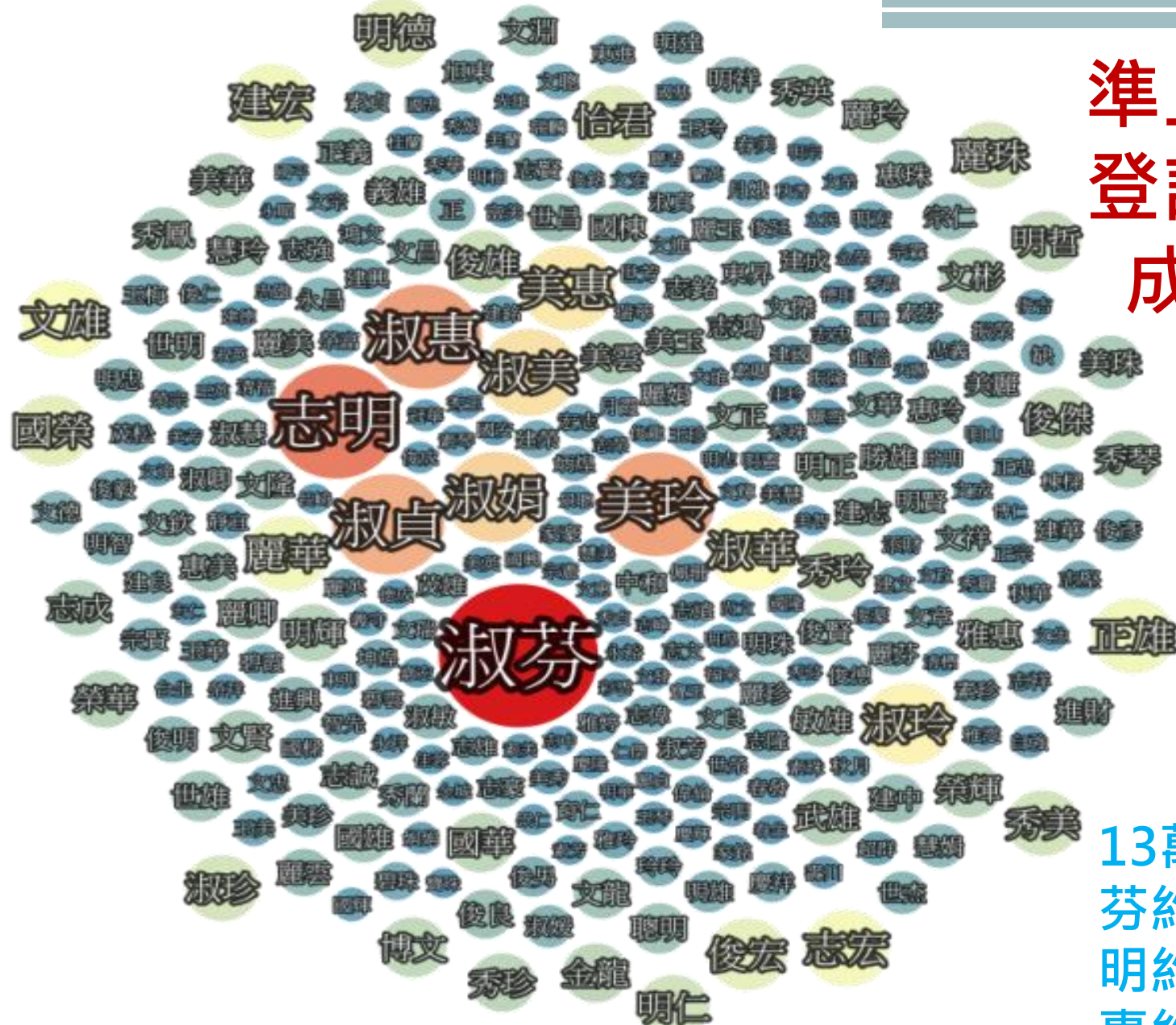
資料清理與篩選

- 公司登記資料約147萬筆
- 使用登記狀況為核准設立、重組、核准報備、核准認許等四類約62萬筆
- 實收資本額登載大於5000萬以上之準上櫃公司條件，則剩約3萬2千家公司

建宏 麗玲 秀蘭
 缺 文龍 玉玲 美芳 正義 美麗 志豪 秀珍 雅惠
 淑敏 志誠 素卿 明宗 進興 志強 秀琴
 俊宏 淑真 文德 美珍 麗雪 世雄 月霞 進益 武雄 麗琴 明輝
 秀梅 玉華 欣怡 玉璇 秋月 建國 秀玉 玉珠 佳玲 靜宜 文英 文章 文祥 榮華 麗美
 素貞 雅玲 建華 聰明 秀惠 俊明 國傑 明正 素娥 佳蓉 文英 志文 美娟 建中 玉梅 進財 文雄
 麗娟 志鴻 明賢 秀華 俊仁 美華 雅華 麗英 月華 素蘭 文生 建吉 國輝 明哲 金龍
 明德 義雄 俊男 啟明 光輝 惠敏 信宏 金淑 淑萍 秀儀 國棟 紫芳 秀卿 惠珍 文賢 建興 金鳳
 慧玲 文正 美蘭 明雄 秋燕 文昌 淑琴 玉珍 建國 鳳興 明志 明智 秀月 永祥 鳳珠 秀蓮 朱仁 榮輝 美玉
 志成 志偉 文欽 美月 家豪 國隆 淑珠 信怡 志鴻 志祥 惠珠 麗芬 淑娟 淑貞 麗珍 慧娟 惠玲 正雄
 麗玉 志銘 秀珠 彩雲 心怡 淑媛 玲玲 秀芬 惠珠 麗芬 淑娟 淑貞 麗珍 慧娟 惠玲 正雄
 明珠 美慧 淑珍 文華 惠貞 明宏 志賢 海卿 美鳳 博文 玉鳳 瑞祥 忠義 世明 文強 福來 俊雄 寶玉
 淑玲 文彬 麗香 秀鳳 碧珠 家銘 素華 俊銘 前前 素琴 雅雯 志雄 富美 麗雲 春美 美秀 麗君 國榮 秀英
 怡君 國華 世昌 俊賢 國雄 美英 淑卿 素珍 明華 世杰 金城 國忠 秀雲 秀娥 志忠 秀英
 麗珠 美珠 淑芳 麗卿 惠美 秀美 美雲 美華

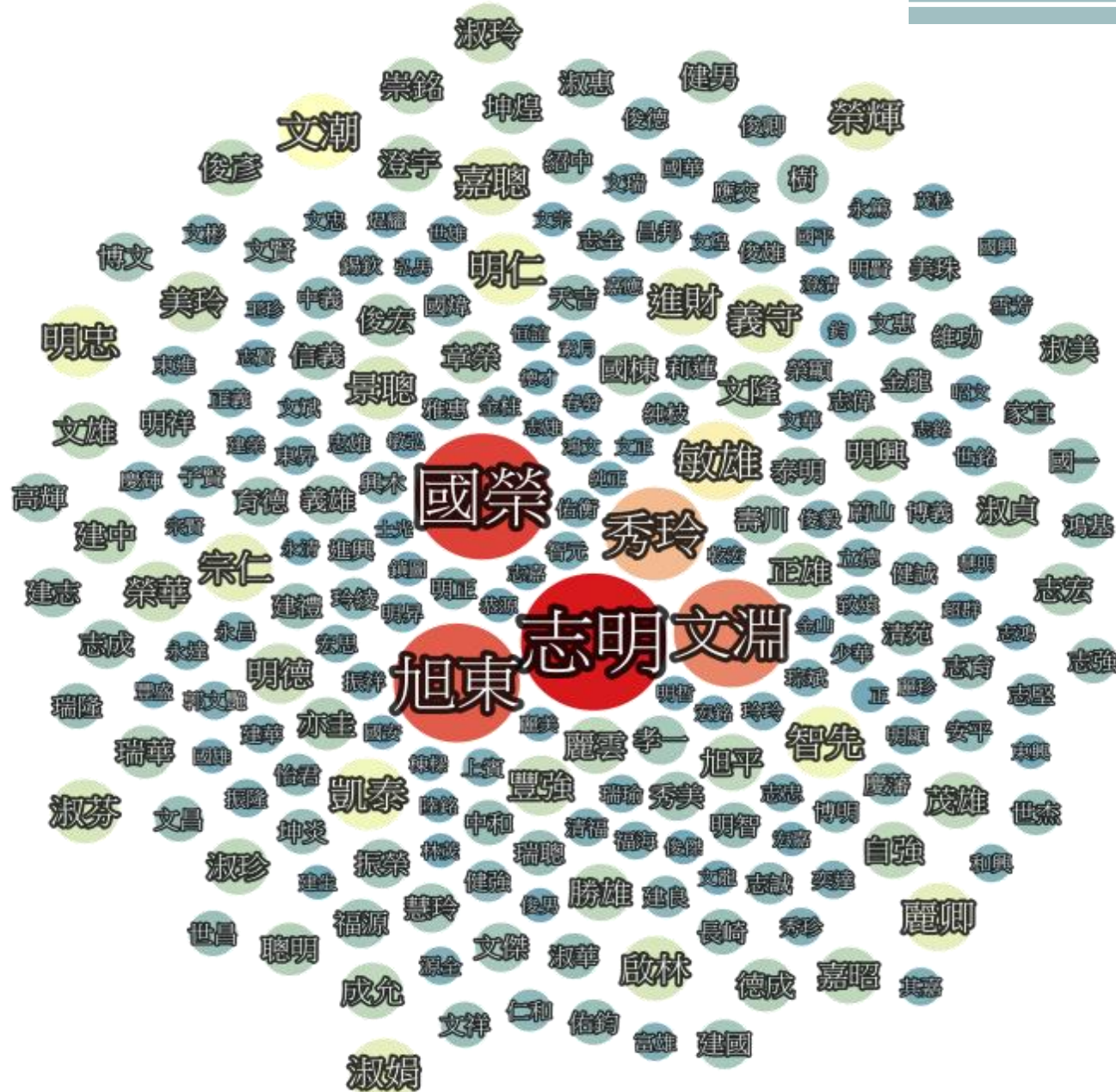
121萬個名字，淑芬約2000次、淑惠約1900次、美玲約1700次

準上櫃公司 登記董事會 成員名字



13萬個名字，淑
芬約220餘次、志
明約170餘次、淑
惠約150餘次

實收資本額 十億以上公 司登記董事 會成員名字



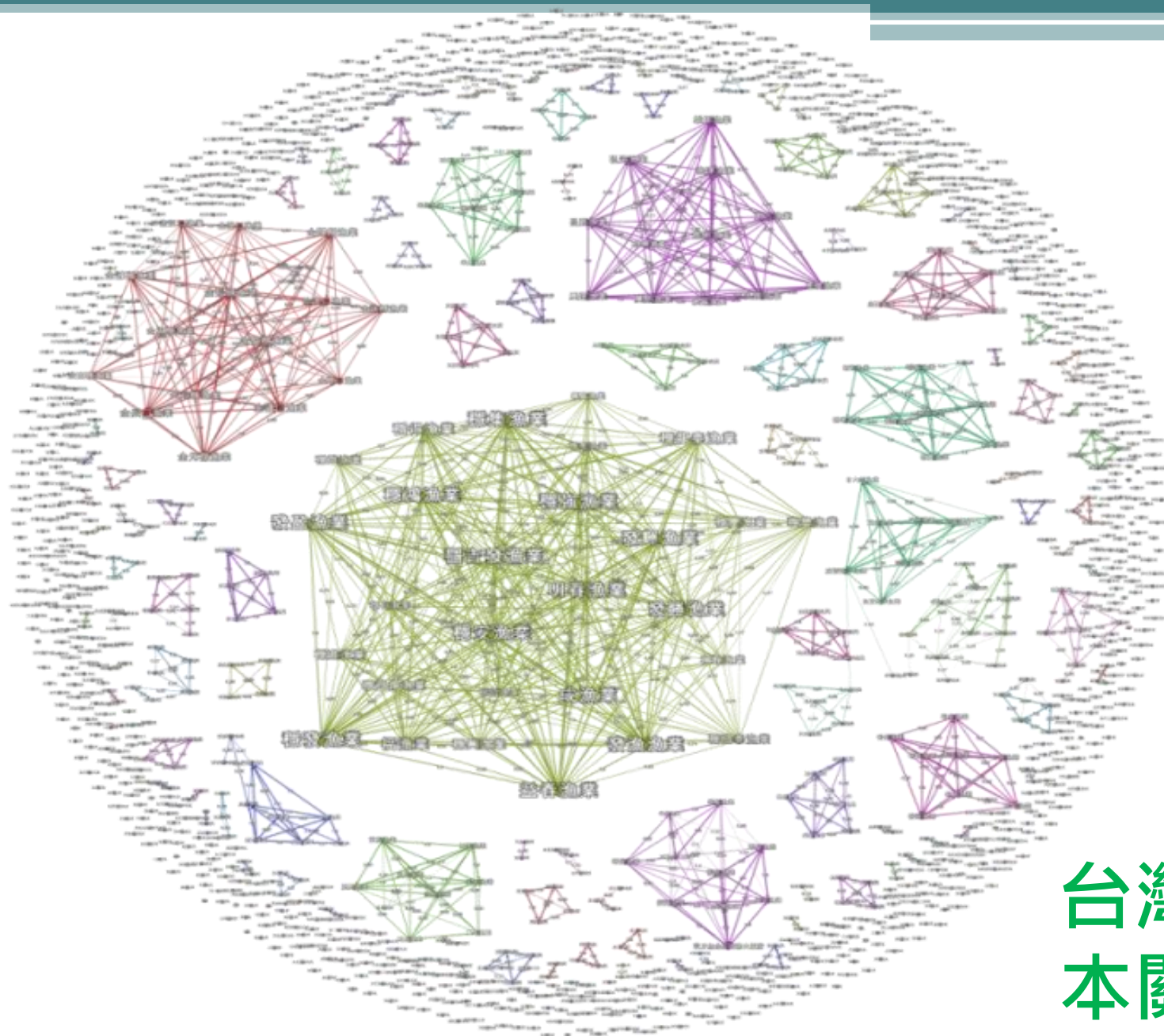
1.6萬個名字，志
明35次、國榮32
次、旭東30次

大數據 + 視覺化技術

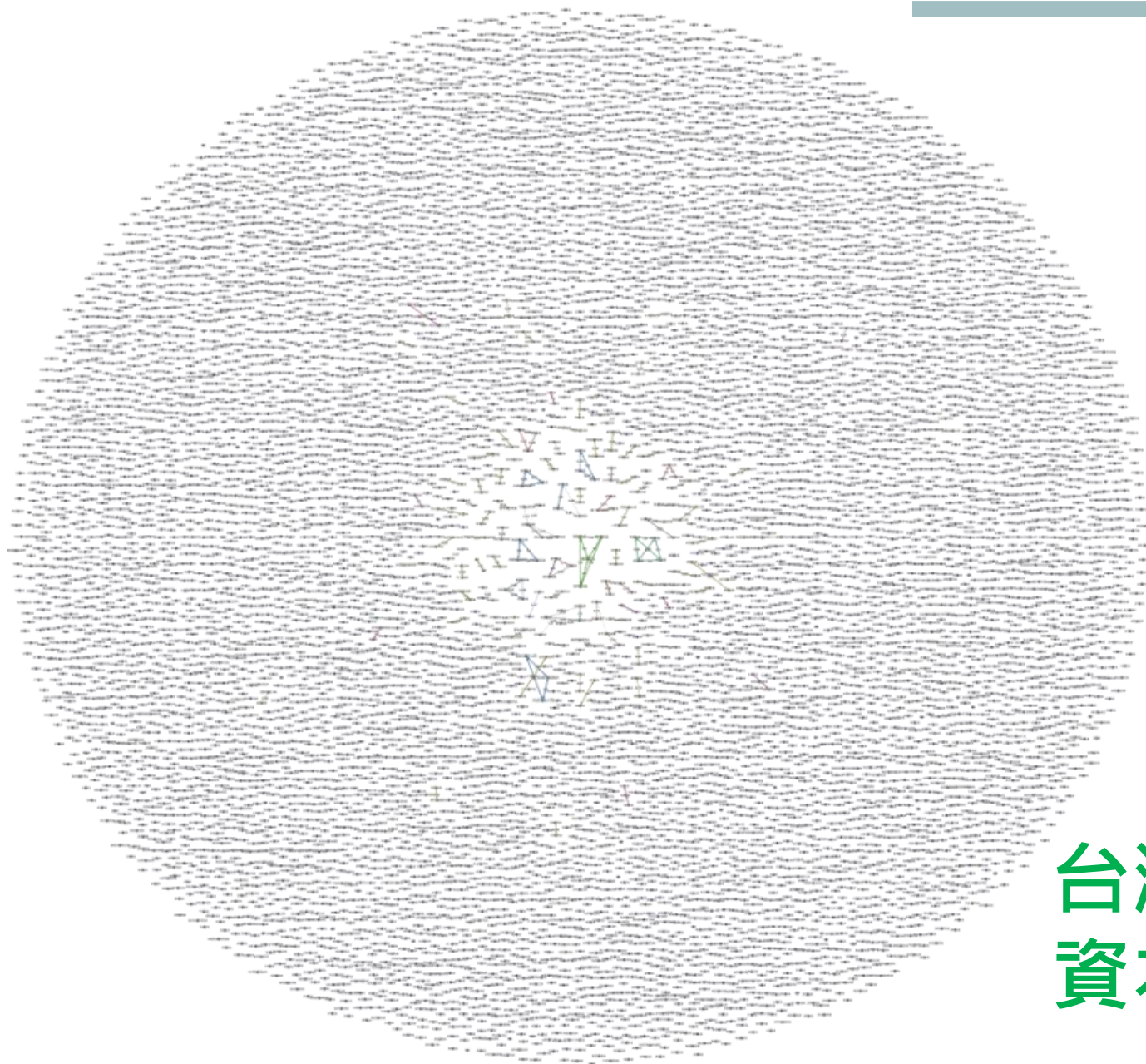
- 顯著分布之常用名字及性別優勢
- 可推測此特定族群的年齡區間
- 隨著公司規模的擴大，性別優勢改變
- 快速浮現顯著資訊及相關差異的簡單範例

公司資本(持股)關係網絡

- 兩家公司之間의 共同董事會成員及相互持股比率
- 一家公司持有另一家公司超過百分之五十以上的股份或資本額，即具有從屬控制關係，稱為關係企業
- 觀察特定產業中關係企業群體及非關係企業群體的分布與結構
 - 以公司名稱後半段之特徵字串及其登記營業項目資料進行產業分類
 - 初步抽樣九個產業：生醫、建設、科技、食品、漁產、工業、旅遊、交通、金融投資



台灣漁業資本關係網絡



台灣食品業 資本關係網絡

準上櫃公司資本關係網絡探討

資本控制集團中心

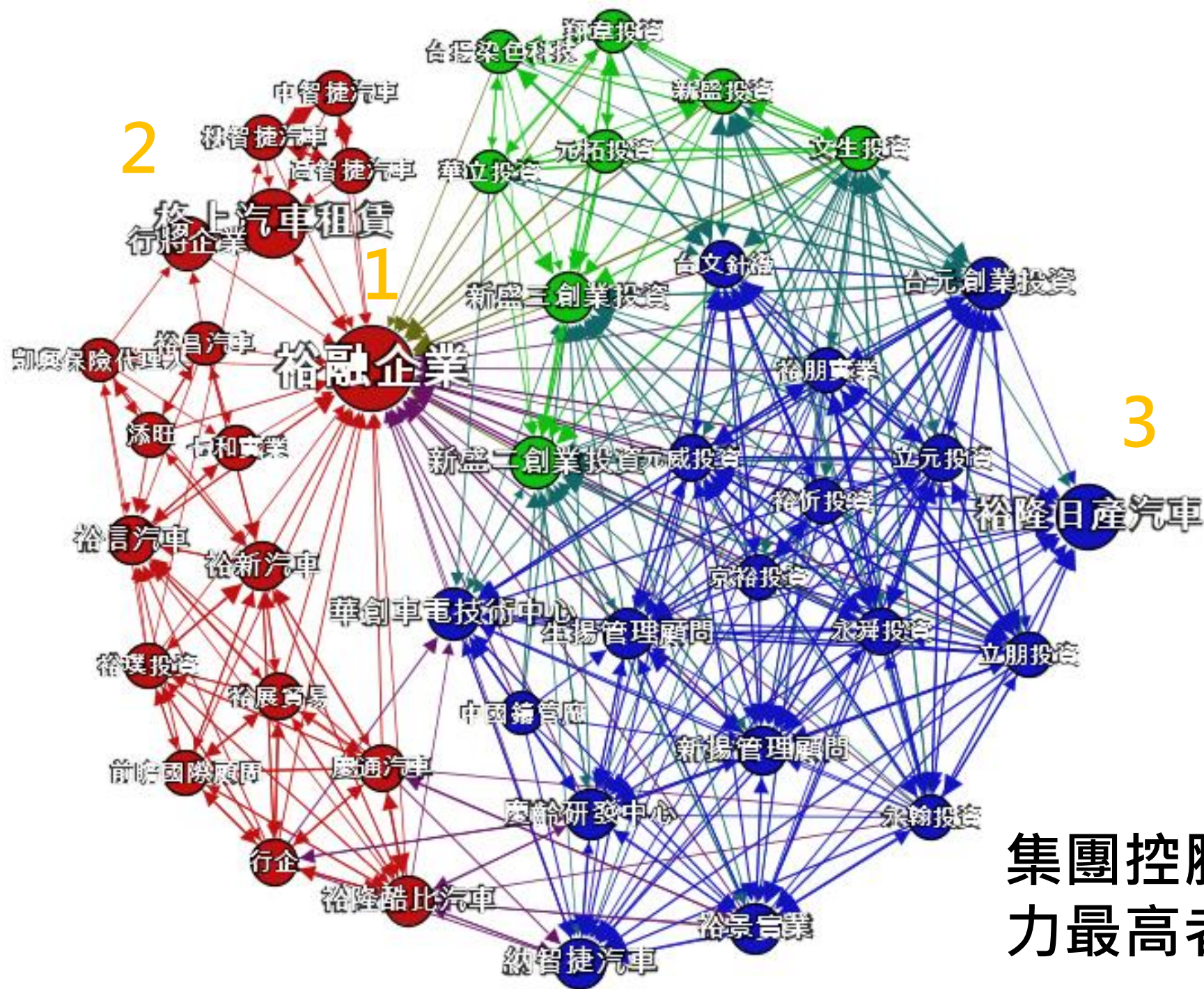
- 第一名：兆豐商銀
- 第二名：裕隆汽車
- 第三名：統一企業



特定企業集團分析範例

企業集團	Node	Edge	Density	Diameter	集中性
裕融企業	44	439	0.232	7	0.79

公司	Page Rank Centrality	公司	Weighted In-Degree Centrality
裕融企業	0.13488	裕融企業	31.36
格上汽車租賃	0.083079	新盛三創業投資	16.26
裕隆日產汽車	0.074741	台元創業投資	15.52
行將企業	0.036735	新盛二創業投資	15.44
新盛二創業投資	0.033278	台文針織	15.03
華創車電技術中心	0.0321	納智捷汽車	14.86
納智捷汽車	0.031718	華創車電技術中心	14.23
新盛三創業投資	0.03003	新揚管理顧問	13.53
裕景實業	0.029935	生揚管理顧問	13.53
台元創業投資	0.029718	裕景實業	12.94

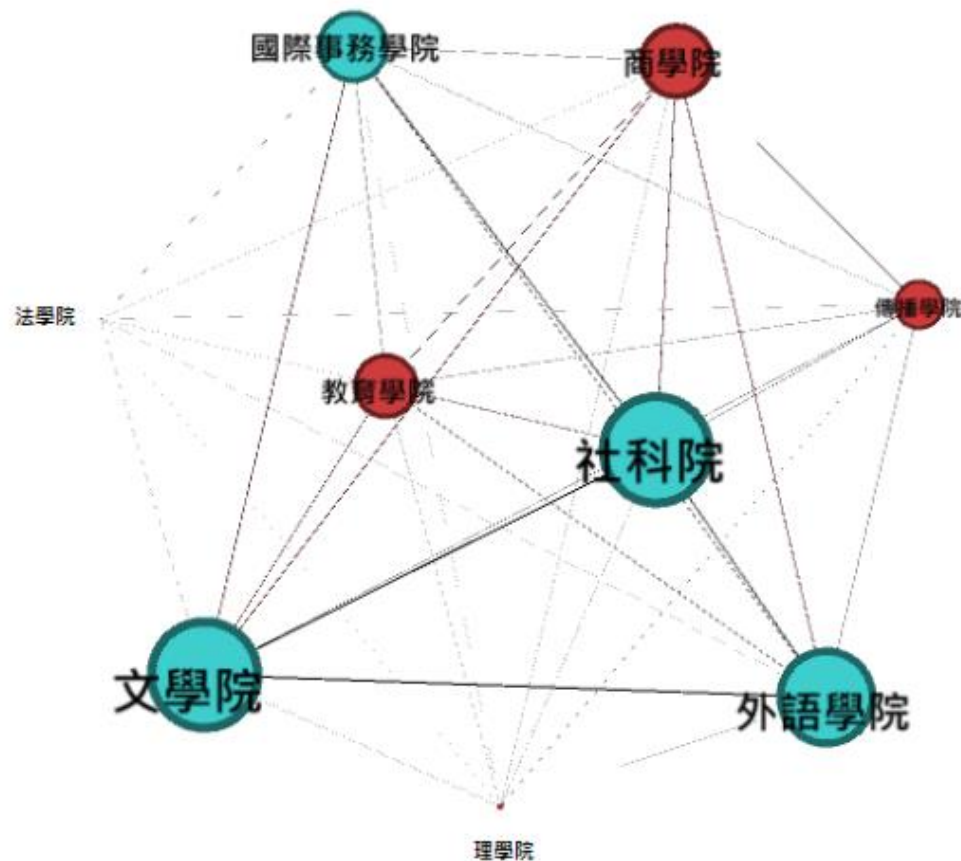


集團控股權
力最高者

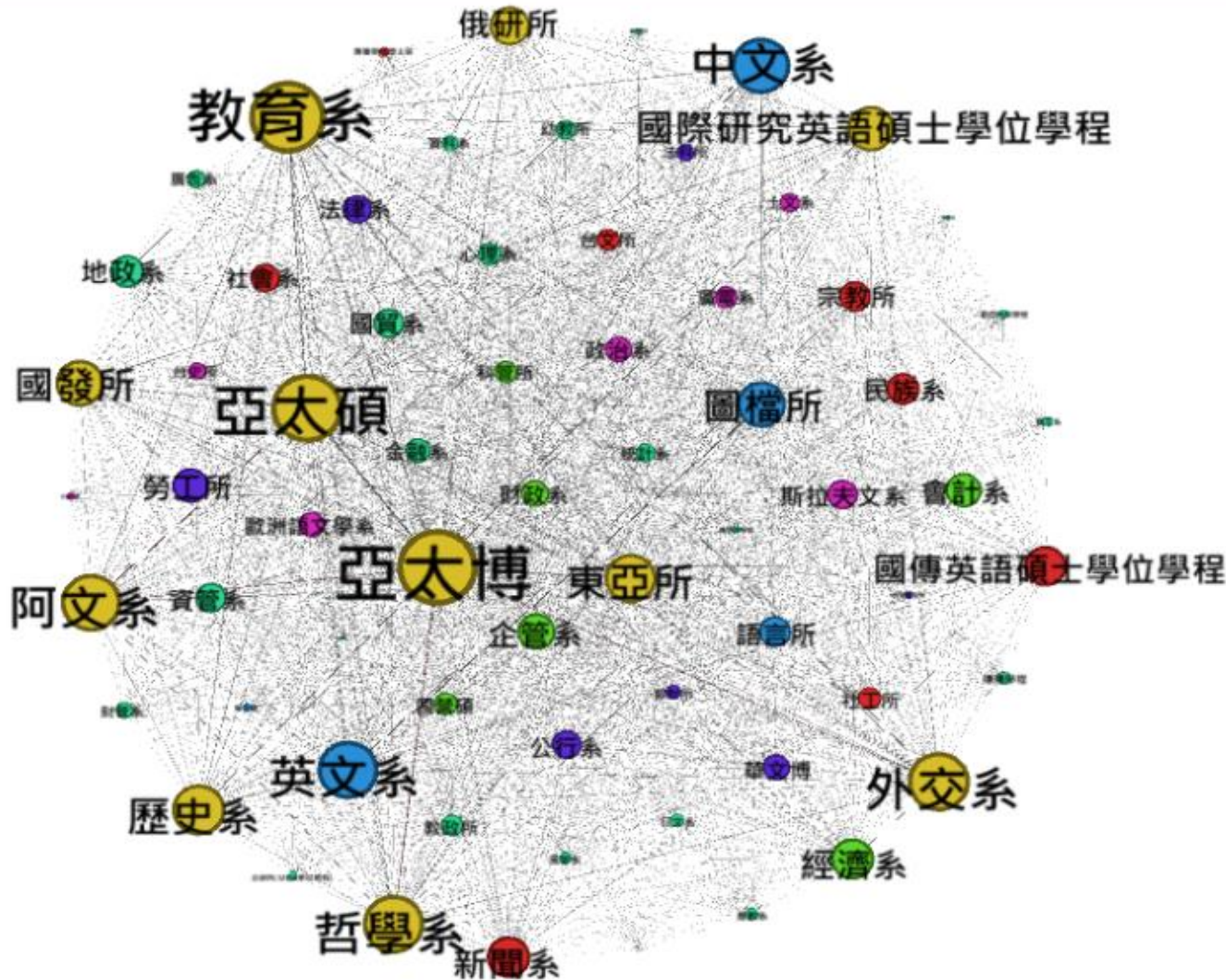
圖書館讀者借閱資料

- 去除個人資料, 只觀察單位資料
- 系所/院 借閱 書籍類別 (十分類/百分類) 的關聯分析

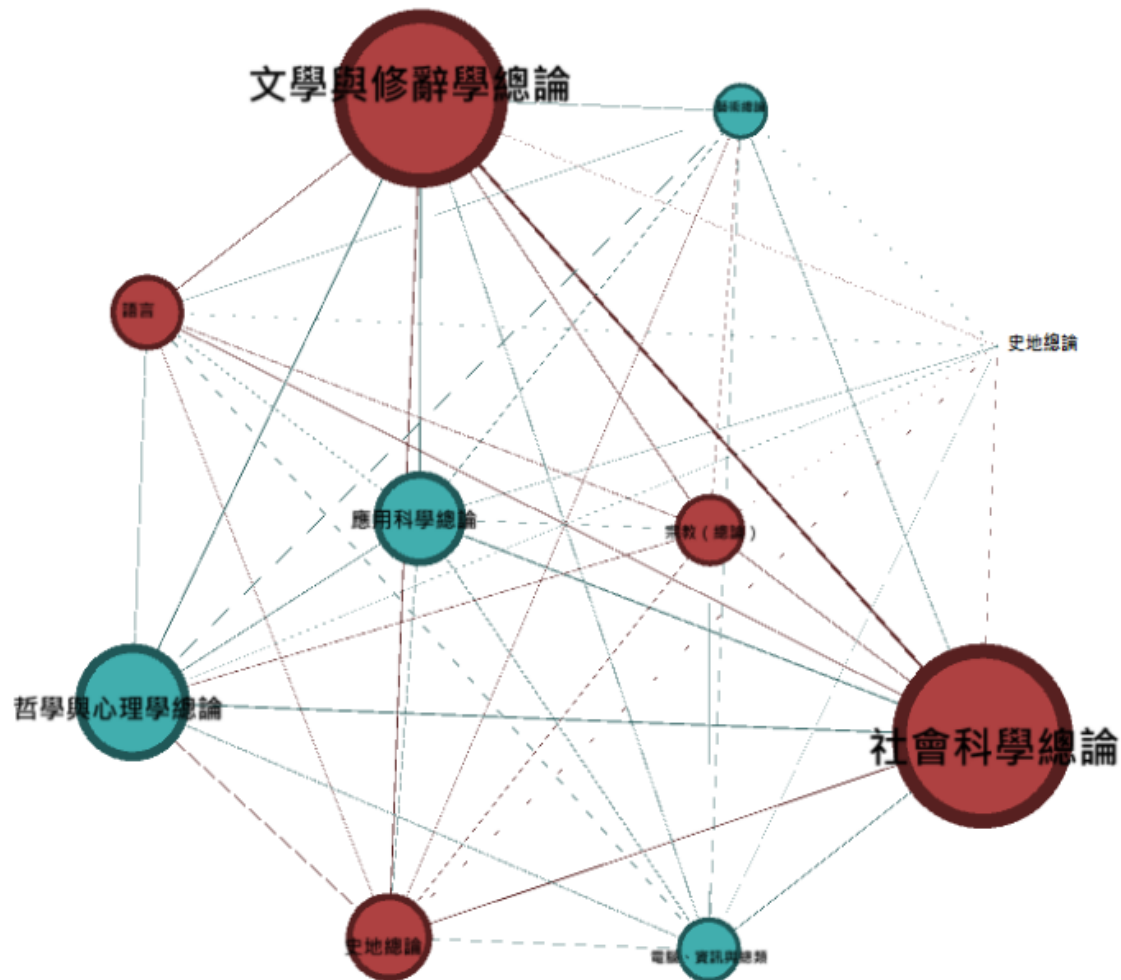
學院之共同借閱關係
(依據十分類)



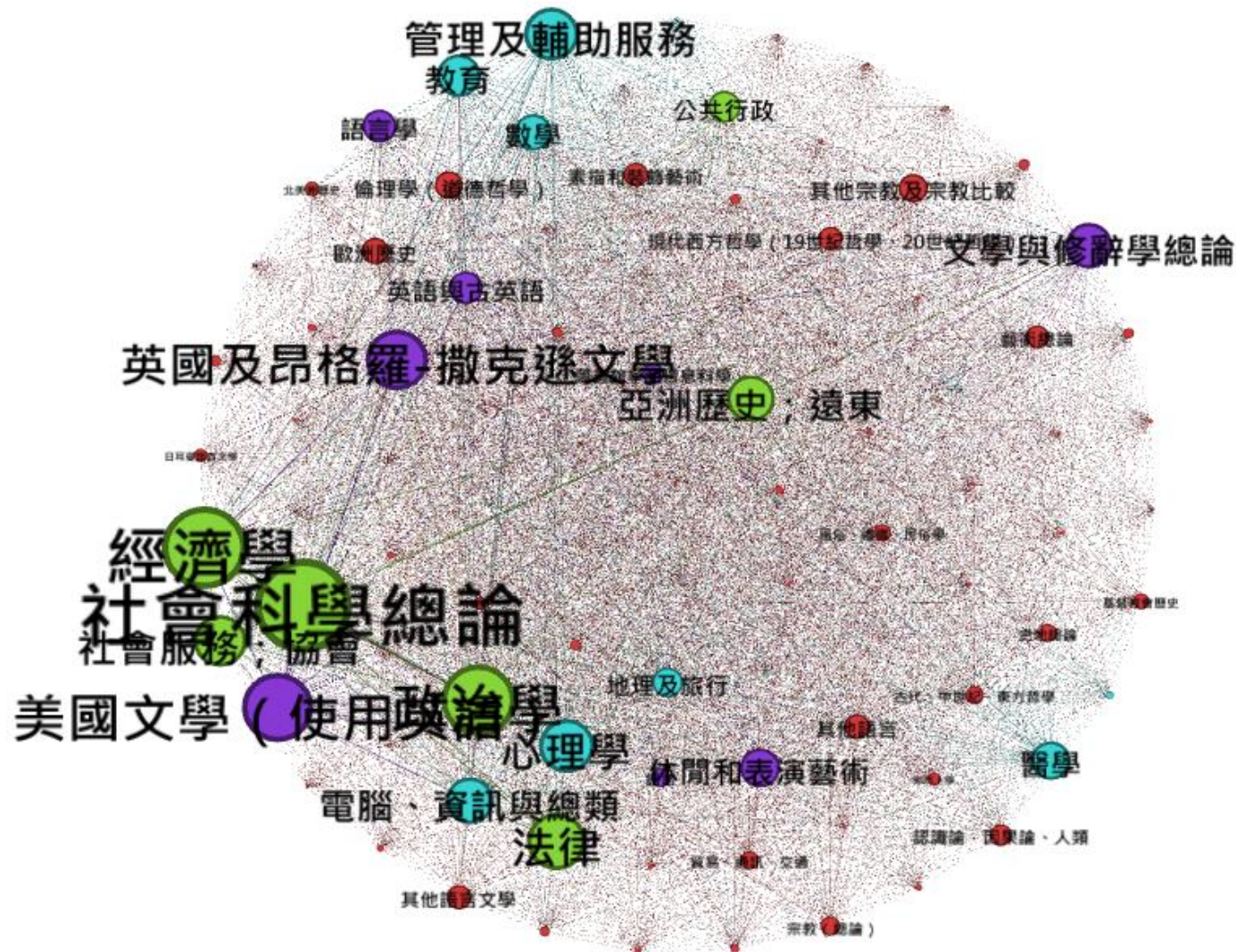
系所之共同借閱關係 (依據百分類)



十分類之共同借閱關係 (依據讀者所屬學院)



百分類之共同借閱關係 (依據讀者所屬系所)



學位論文資料應用範例

- Tinayu Ying and Honggen Xiao, Knowledge Networks – A Social Network Analysis of Dissertation Subjects
- 資料: ProQuest Dissertations and Theses – Full Text (1994-2008)
- 主題: tourism
- 檢索詞: travel, traveler, tourism, tourist, ecotourism
- Metadata search: keywords and title

觀察某一特定領域的知識發展脈絡與過程

Table 1 Longitudinal Subject Network Measures

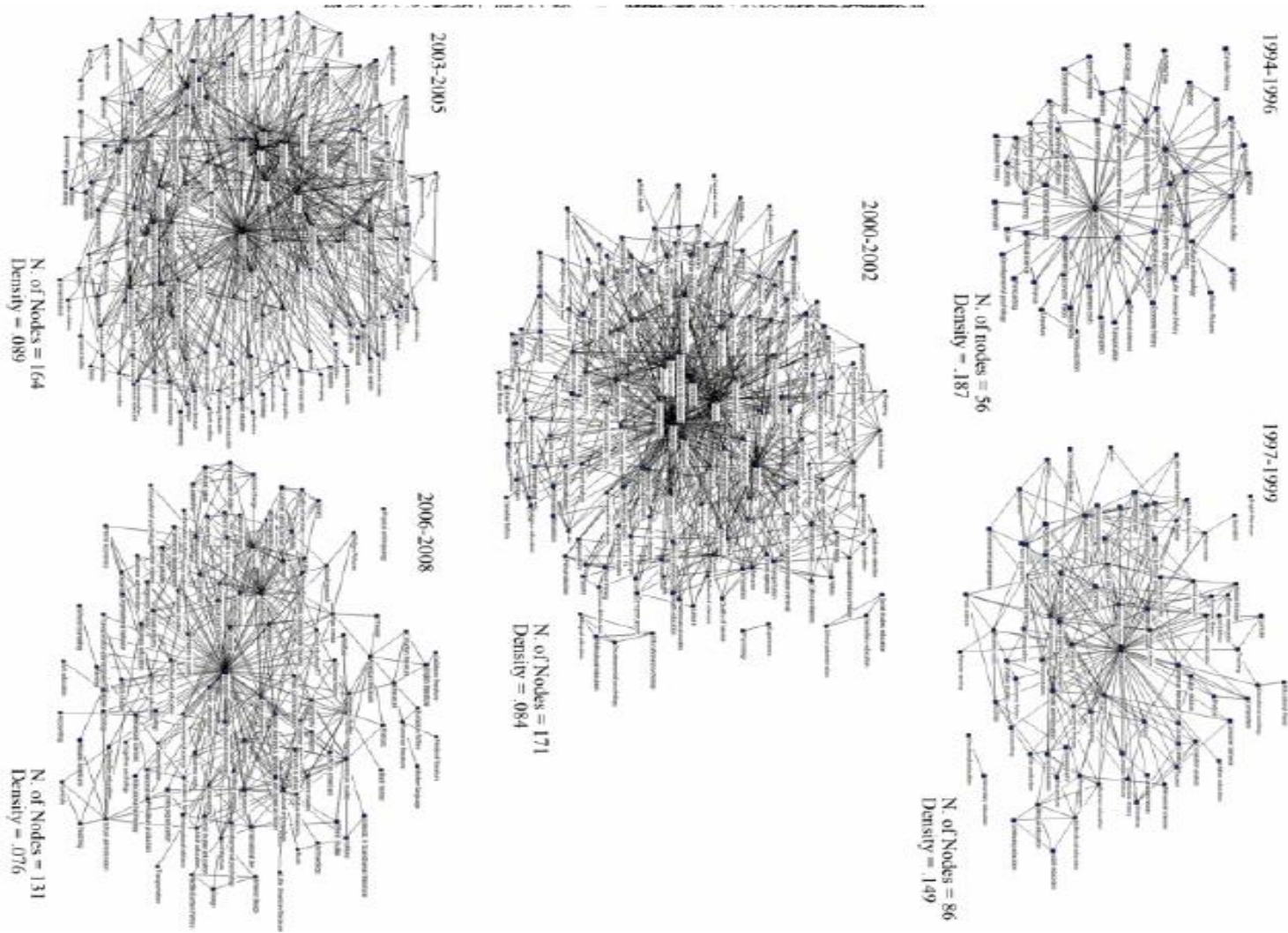
Time Period	Overall	1994-96	1997-99	2000-02	2003-05	2006-08
No. of Dissertations	812	101	130	168	188	225
Network Size	304	56	86	171	164	131
No. of Ties	3650	290	586	1552	1584	852
Network Density	.0396	.0942	.0802	.0534	.0593	.05
Betwn. Central. Index	52.75%	67.63%	64.63%	49.77%	48.67%	65.48%

基本量化觀察

- 論文篇數從1994年的31篇, 到2008年的90篇, 幾乎成長3倍
- 每篇論文的關鍵詞數量, 從1到10, 平均3.14
 - 關鍵詞數的分布
 - 關鍵詞平均數量的歷年變化

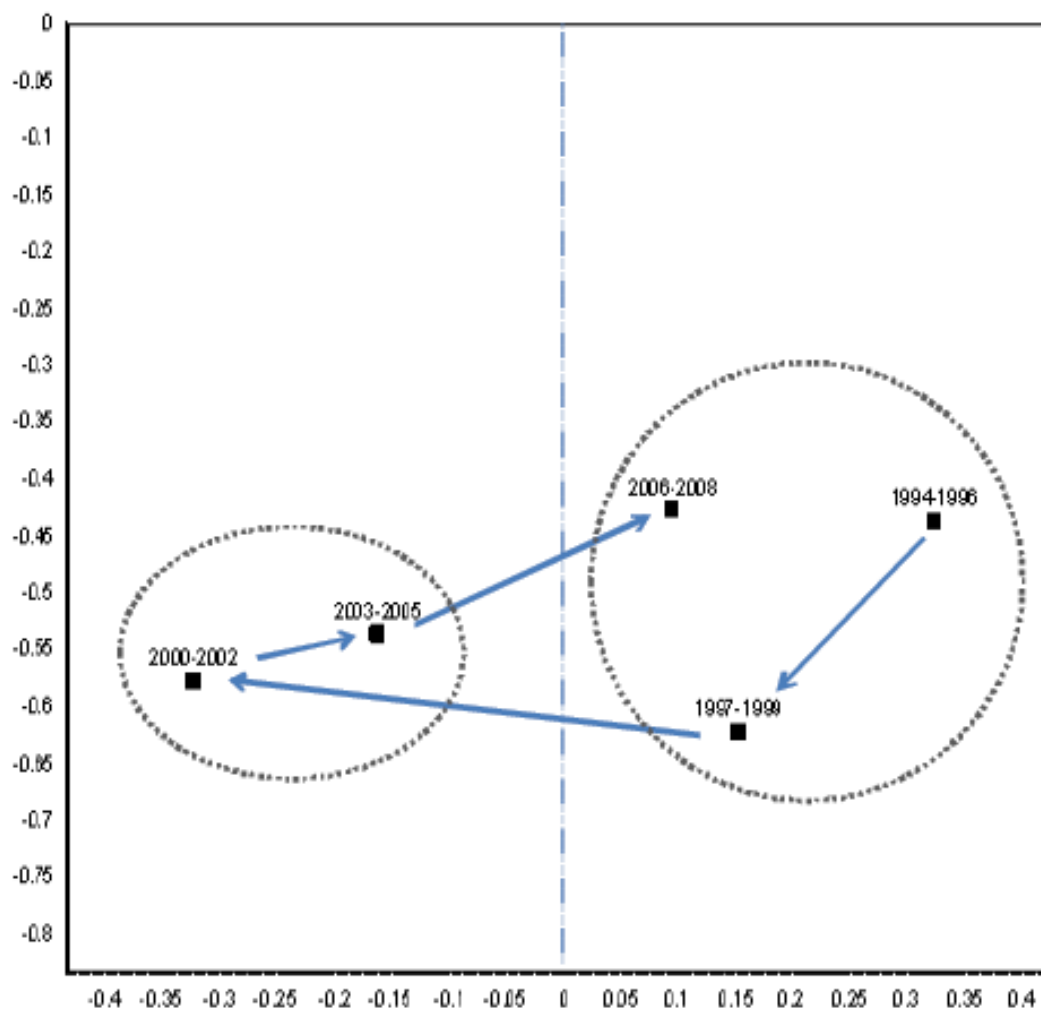
觀察某一特定領域的知識發展脈絡與過程

五個年度期間的主題網絡



五個主題網絡的結構關聯分析

Figure 3 MDS of Five Periodical Networks

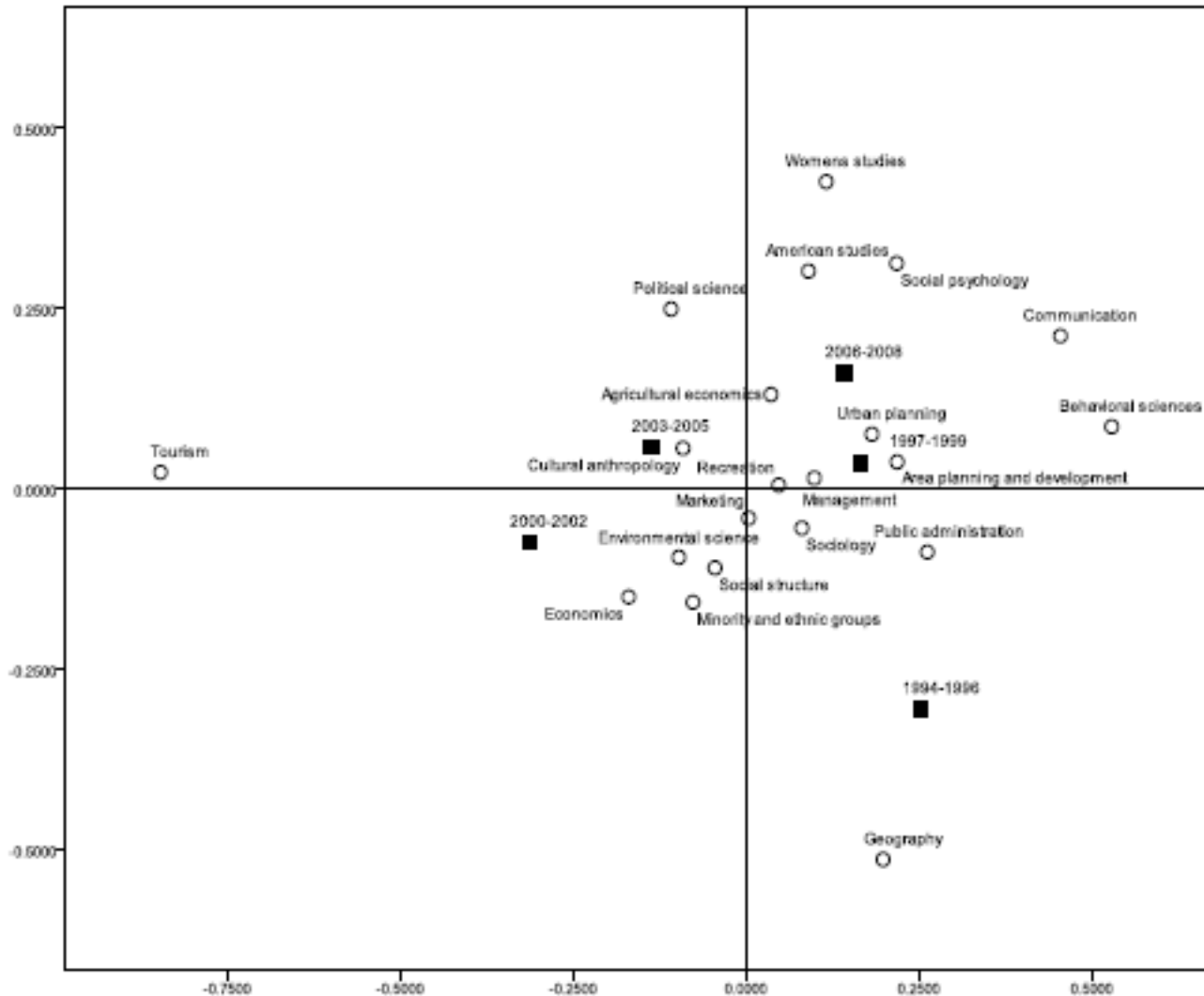


以核心/邊緣模型 (core/peripheral model) 從304個關鍵詞抽取出21個主要關鍵詞

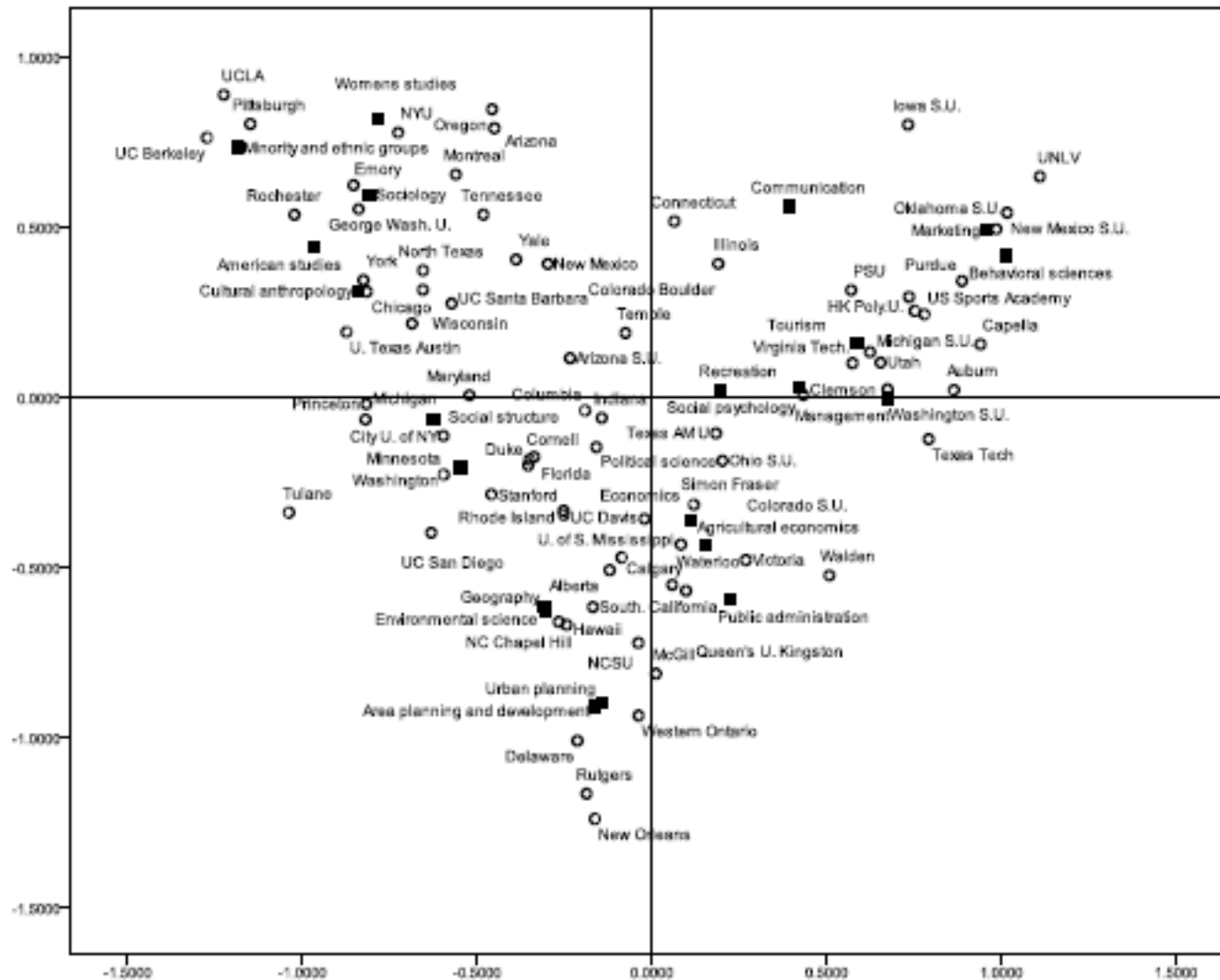
Table 4 Ego-network Measures for the 21 Subjects in Overall Network

Subject	Overall Size	Density	N. Ego. Betwns	Periodical Egonetwork Size					Period. Egonetwork Tie Changes				
				94-96	97-99	00-02	03-05	06-08	T1	T2	T3	T4	Mean
Recreation	249	4.55	61.22	42	69	141	129	90	47	108	138	123	104
Tourism	157	7.59	38.72	0	7	87	87	34	6	84	110	89	72.25
Marketing	135	8.46	35.33	12	23	78	61	36	23	73	85	71	63
Economics	85	16.42	31.41	6	13	47	39	13	11	50	62	42	41.5
Management	75	14.38	35.12	8	9	27	42	18	11	30	49	48	34.5
Cult. anthropology	63	21.52	32.54	11	28	33	30	26	21	23	33	28	26.25
Urban planning	54	27.95	15.52	8	19	24	31	20	13	21	31	31	24
Area plan. & dvlpt	53	28.16	15.48	8	19	24	31	18	13	21	31	29	23.5
Envir. science	51	24.16	28.05	7	16	31	28	6	13	31	31	24	24.75
Geography	45	26.36	26.79	22	13	21	20	9	17	20	21	19	19.25
American studies	42	28.34	21.76	11	18	16	19	18	17	18	19	17	17.75
Sociology	40	32.56	17.85	13	15	9	25	16	10	12	18	23	15.75
Social structure	38	43.24	10.63	11	16	22	17	8	15	16	23	17	17.75
Agri. economics	35	28.75	26.9	5	8	19	12	6	9	21	25	12	16.74

時間與研究主題的連結關係



學校機構與研究主題的連結關係



大數據分析與數位人文意涵

- 過去人文社會資料限制
 - 資料覆蓋性
 - 資料處理成本
 - 人腦對資料接收與解讀之認知能力
- 鋪天蓋地的數位浪潮帶來巨大的衝擊
- 資料原生性 (+ 回溯數位化)
 - 人類歷史與現代樣貌某種程度的被投射到一個數位空間

大數據分析與數位人文意涵

- 資料可操作性

- 數位本質讓各種計算技術有非常寬廣的操作空間
- 認識並理解各種數位資料操作概念與技術，有助於產業創新應用

- 資料可連結性

- 各種資料之間在主題、對象或概念上可能存在連結，而能提供更完整的意義與樣貌

大數據分析與數位人文意涵

- 資料可呈現性

- 資料的維度與面向可以被輕易的選擇、調控與投射，以多視角的呈現資料內容的豐富樣貌
- 資料視覺化技術更以多元圖像呈現資訊意涵，展現出更有力的資訊傳遞效果

- 資料可檢驗性

- 資料就如同實驗室中的材料，可以反覆操作測試，實驗過程與結果都可以被複製檢驗，而提升研究成果的客觀性與正確性

大數據分析與數位人文意涵

- 資料高價值性

- 不斷生產堆積的資料大幅度的捕捉了人類行為與社會運作的真實樣貌，提供了偵測、探究、理解個人與群體的全新途徑
- 數位資料礦源不論是對學術研究、政府公共治理、知識經濟產業，都是具有高度價值而應積極掌握利用的重要資產

大數據資料分析優勢

- 可以宏觀/縱觀, 也可以聚焦微觀



- 可以量化分析, 客觀評估(+採樣優勢), 可以區別顯著, 可以浮現, 可以非線性連結

大數據功能意涵

- 協助人類拼湊出事物的接近全貌, 建構對特定事物更真實的了解



大數據功能意涵

- 協助建立能描述特定事物的**模型**, 歸納各種**型態****機制**, 取得對情勢變化的適切(與快速)反應能力



大數據功能意涵

- 協助將電腦對特定事物的**判斷決策**能力調理教導 / 調整校正到更高層次, 趨向於更正確, 更快速反應, 甚至超越人類能力 (受限於認知能耐)



對圖書資料的新意涵

- 建立資料的新意識

- 過去的資料: 紙本, 平面, 線性, 少量
- 過去的資料利用: 長期發展的知識框架下, 人力整理, 編輯, 搜尋, 人的閱讀理解與詮釋
- 現在的資料: 數位, 多維度, 可切割, 大量
- 現在的資料利用: 更廣更深的電腦輔助與介入

- 認識計算思維(computational thinking)

- 資料 + 計算 (想法, 邏輯)

對圖書資料的新意涵

- 掌握對資料內涵的既有優勢, 但應適時與數位時代的新思維, 新框架介接, 演化

Data Curation <i>(RDM)</i>	Data Resourcing <i>(subject librarian)</i>
Data Reporting <i>(internal audit/ analytics)</i>	Data Sensemaking <i>("data literacy"/ infoskills)</i>

資料等級

- Level 0: 參雜許多錯誤之原始資料
- Level 1: 去除明顯錯誤之原始資料集
- Level 2: 初步整理之資料集, 如標準化、編碼等
- Level 3: 整合性資料資源, 連結相關資料集
- Level 4: 經人工標註、分類、驗證, 可直接進行分析、檢驗、探索, 並可提供學術社群共用

(Reference: Digging into Data Challenge)

圖書館的角色

- 數位人文學術基礎設施, 包括資源、工具、空間、設備等的健全打造
- 數位人文發展平台
 - 跨單位, 跨領域, 跨層級的中介連結
 - 介於學術與行政之間
 - 延續性與服務性

圖書館的自我意義

- 直接參與推動一個新興學術領域的發展歷程
 - 過去是提供參考資源的間接旁觀者
 - 轉換為肩負重要推手角色
- 圖書館領域知識與功能進化的任務目標
 - 資源建置, 資源描述與連結, 資源加值利用...
- 館員專業角色的提升, 圖書館展現為充滿活力與不斷自我提升的有機組織

新世代圖書館員的專業與技能

資料管理師

資料設計師

資料分析師

資料增值師

.....



政大數位資料稜鏡實驗室
Digital Lab for Data Prism