巨量資料中的小世界 -- 漫談社群網路

李政德／臺大資訊網路與多媒體研究所碩士班
林守德／臺大資工系教授

一、引言

隨著網際網路的無所不在，以及上網速度的提升，Facebook、Twitter 與 LinkedIn 等線上社群服務（Online Social Service）迅速竄紅。在這些線上社群網站所提供的服務，除了能夠幫助使用者認識新朋友、找回舊朋友以經營管理自己的人際關係，更可透過「按讚」、「分享」與「評論」等功能對感興趣的各種資訊與朋友們進行互動。此外，再加上人手一機的智慧型手機時代，使得人們與這些線上社群服務的關係變得密不可分。眾所周知，線上社群服務最根本的核心是一個連接世界上數以億計使用者的社群網路（Social Network），這個存在於你我之間無形的社群網路究竟長得是什麼樣子呢？社群網路其實是一門橫跨資訊科學、物理學、社會學、生物學、管理學等領域的學門。本文將給各位一個關於社群網路分析的簡介，從何謂社群網路及其歷史、不同領域研究社群網路的觀點，到當今社群網路分析研究幾個重要的課題。

什麼是社群網路

最原始的社群網路是一種由節點（node）與邊（edge）所組成的圖形結構（graph），其中節點所代表的是人，邊所代表的是人與人之間的各種相互認識的關係。通常社群網路的建構來自一特定資料集（corpus），例如 Facebook 社群網站、公司組織內部專家網路、社團內部人際關係。由資料集所建構的社群網路並非全然靜態（static），隨著時間會有新的成員加入，節點數因而增加，同時新成員與舊成員彼此間會彼此認識，進而產生新的邊，我們將這種隨時間演變的圖形結構稱為「動態社群網路」（Dynamic Social Network）。此外，社群網路中的個體會有不同的身分類型（如學生、老師、演員與導演等）或屬性（如性別、興趣與專長等），且個體間的連結關係也有非常多種可能類型，如朋友關係、家人關係、師生關係與合作關係等，這種考慮節點與邊之類型的圖形結構稱為「異質性社群網路」（Heterogeneous Social Network）。由上

可知，社群網路視為一種能夠有效表示個體間各種互動與關連的資料結構，有別於許多資料集是獨立且來自相同機率分布（Independent and Identically Distributed, IID），社群網路透過連結進一步允許個體間具有多重關聯性（relational），且這些資料彼此間的關聯性是會隨時間演進的（time-evolving），因此更能描述真實世界複雜的個體互動。而社群網路分析（Social Network Analysis, SNA）正是研究這種複雜網路結構的一個學門。

社群網路分析的歷史：從微量資料（Small Data）到巨量資料（Big Data）

1960 至 1990 年代，由社會學家率先從事社群網路結構與角色的研究，當時社群網路資料的來源多透過人工蒐集，藉由發送問卷到一特定團體（如學校、公司或社團），詢問每個調查對象認識哪些朋友，將每個受試者所認定的幾十筆交友情形之微量資料視為社群網路的一小部分，接著把受試者的社群網路蒐集整合，以紙筆畫出社群網路結構，並用肉眼進行與質性方法分析。因此，可以想見這種人工方式所蒐集建構出來的社群網路不僅資料量少，所描述的社交結構也僅限於區域性的，難以蒐集到完整的朋友資訊，加上紙筆所能描繪的網路大小有所限制，通常至多僅能分析到上百個節點與邊的小型社群網路，因此分析的結果也僅限解釋特定群體的社交互動，無法代表真實世界完整的社群網路。

從 2000 年代至今，隨著 Facebook 與 Twitter 等社群網站的興起與竄紅，使用者數不斷攀升，截至 2014 年三月，Facebook 每天有 8 億個活躍的使用者，可以想見這些使用者背後是一個節點與邊均數以億計的巨型社群網路。這些線上社群網路多提供應用程式介面（Application Programming Interface, API）讓開發者能夠在獲得充分授權下撰寫爬蟲程式（crawler）來取得巨型社群網路圖形資料（Big Graph）。由線上社群服務所建構而得知巨量社群網路之社交資料橫跨學校、政府機關、國界等不同群體，能夠同時描述並結合每一位使用者在不同社交圈的人脈網路，進而形成連接全世界的社群網路，這種資料的完整性是傳統人工作法所無法達到的；此外，線上社群服務要求朋友關係必須透過雙方同意才能建立，因此相較傳統人工田野調查所建構之社群網路，其巨型社群網路分析之結果可信度較高。另一方面，資料探勘（Data Mining）與大數據（Big Data）相關技術的發展，譬如 Google 與 Apache 所開發設計的 Hadoop/MapReduce 分散式資料平行運算處理架構，以及 NoSQL 資料庫的發明，甚至 Google 近年專為圖形演算法所開發的 Pregel 架構，都讓研究人員能有效率地對巨型社群網路資料進行儲存與分析。

一個跨領域的學門

社群網路分析乍聽之下會讓人以為是社會科學的研究課題，其實它是一橫跨社會學、物理學、資訊科學、生物學、管理學、經濟學與心理學等領域的學門。不同領域的研究人員從不同角度切入探討社群網路的不同議題，或者利用社群網路來輔助各自領域的研究。以下我們以社會學、物理學與資訊科學為例，簡介不同領域對於社群網路分析的思維以及幾項重要研究課題。

社會學研究社群網路主要在於透過質性方法，分析節點與邊在社群網路扮演的角色，以及比較個體間不同互動結構對於整體社群網路之影響。計算節點之中心程度（centrality），即位居一個社群網路的中心位置，是最基礎的課題，中心節點普遍被認為在社群網路中扮演重要角色，其中最重要的三種中心度指標為「度中心性」（degree），定義擁有愈多鄰居的節點為中心節點；「親近中心性」（closeness），定義與其他節點之平均距離越短的節點為中心；「間接中心性」（betweenness），定義頻繁地被其他節點間之最短路徑經過的節點為中心。角色分析（Role Analysis）目的在於識別社群網路中扮演相同角色之節點，如哪些節點是扮演父母親的角色、或哪些節點是公司主管的角色。結構平衡理論

（Structure Balance）假定社群網路每個邊上有一個正或負的標籤，正代表彼此是朋友，負則代表彼此是敵人，並指出在以三個節點為基礎的三角形結構中，具有三個正邊或一個正邊兩個負邊之三角形才算是結構平衡，即三個人之中，彼此互為朋友，或者兩個朋友並有一共同敵人，個體間才不會有競爭關係。

另一方面，社會心理學家 Stanley Milgram 做了一個對社群網路研究影響深遠的「小世界實驗」（Small-world Experiment），其目的在於探討社群網路究竟有多大（或多小）。其透過信件傳遞的實驗發現，連結世界上任意兩個人平均而言只需在社群網路中走六步即可到達，即平均只需要五個中間人就可以聯繫任兩個互不相識的人，是著名的「六度分離」（Six-degree of Separation）。此外，社會學家 Mark Granovetter 提出「弱連結」（weak tie）的概念來解釋六度分離，他認為通常與我們生活圈接近、互動頻繁的人多因物以類聚而同質性高，所獲取的資訊也較為相近，這種朋友關係屬於「強連結」；我們可透過與某些較少來往的泛泛之交進行互動，他們能提供不同資訊，帶來另一個圈子的知識和機會，這種與泛泛之交的連結稱為「弱連結」。這些弱連結雖然為數不高，但是卻具有連結不同族群的能力，進而造成小世界的現象。

物理學對於社群網路分析的研究思維，主要觀察真實世界各種社群網路的具有什麼共通的特性與現象，並設計理論數學生成模型（generative model），無中生有產生具有這些特性的社群網路，來解釋真實世界複雜的社群網路。理論物理學家 Mark Newman，觀察與歸納出十幾種重要的社群網路特性，其中最重要的三個為：(1) 低平均路徑距離（low average path length）：任兩節點之平均距離是短的，如上述的六度分離現象，代表資訊能在網路中能迅速傳播；(2) 高群聚係數（high clustering coefficient）：即一個人的朋友們傾向也彼此成為朋友，形成三角形結構，網路中愈多三角形，則群聚係數愈高；(3) 節點度冪次分布（power-law degree distribution）：即社群網路中朋友數很少的人相當多，朋友數非常多的人相當少，且呈現指數級遞減。

物理學家 Duncan Watts 與數學家 Steve Strogatz 於 1998 年提出「小世界模型」（small-world model），該模型認為只要隨機將格點圖（ring lattice）（即點與邊如棋盤狀規則排列之圖形）中少數節點的邊，重新連接（random rewiring）到任意其他節點，即能生成兼具低平均路徑距離與高群聚係數的社群網路。物理學家 Albert-László Barabási 與 Réka Albert 於 1999 年提出「無尺度網路模型」（scale-free model），該模型假設新成員是逐一被加入社群網路的，每當有新節點加入，它的連結生成方式遵從「偏好依附原則」（preferential attachment）：有較高的機率連接到現有社群網路中朋友數較多的節點，以此方式即可生成出具有節點度冪次分布之社群網路。

資訊科學家對於社群網路分析之研究主要有三大方向：圖形理論（Graph Theory）、機器學習（Machine Learning）與自然語言處理（Natural Language Processing）。圖形理論的精神在於透過圖形演算法之設計，在以較短的時間以及較少的儲存空間，計算量測各種網路之特性與指標、或尋找圖形中各種定義之特殊結構。例如在一社群網路中找出任二節點之最短路徑，用以計算一個社群網路的大小；在一社群網路中找出內部節點彼此緊密連接、與外部節點連接較為鬆散之網路社群（communities），代表著一群群物以類聚、興趣相投的人；偵測社群網路中扮演關節角色之節點（articulation vertex）或扮演橋樑角色之邊（bridge），若他們被移除，網路將變得四分五裂；尋找社群網路中代表個體經常彼此互動之行為的頻繁子圖樣式（frequent subgraph patterns），例如了解社群網路具有哪些技能的人及其彼此之合作模式會有較高的效率產出；或偵測個體於社群網路中之特殊行為（anomaly patterns）。

機器學習的主要精神在於從現有資料中學習並建構出捕捉資料關聯性規則的模型（model），並利用學習而得之模型對未知資料進行預測（prediction），例如從購買商品的瀏覽或歷史紀錄中學習使用者的喜好，進而預測推薦新的商品給使用者。機器學習運用於社群網路分析，將能學習出每個人交朋友的規則，進而預測或推薦未來你可能會認識的朋友，此課題稱為連結預測（Link Prediction），譬如多數人有較高的機率去認識朋友的朋友，或者傾向於認識與自己興趣接近的人交朋友，又或者有些人喜歡認識朋友數較多的名人；機器學習也可以幫助預測社群網路中每個人的興趣、專長與其他個人屬性（如性別、年齡與居住地），此課題稱為節點標籤預測（Node Label Prediction），其背後原理是物以類聚（homophily）的概念，具有類似興趣或屬性的人會有較高的機會彼此認識與進行互動；此外，我們也可以透過機器學習將社群網路分析與商品推薦（Item Recommendation）結合，利用口碑行銷（Word-of-Mouth）的原理，將好友所喜愛的商品推薦給使用者，此種社交推薦方式（social filtering）已被證實更為有效，相較於傳統根據商品相似度來做推薦之內容過濾（content filtering），以及根據使用者喜好之相似度來做推薦之協同過濾（collaborative filtering）。

自然語言處理之目的在於設計計算模型，讓電腦能自動分析並理解人類所使用的各種文字語言之結構與內涵，並作進一步的應用如機器翻譯與資訊檢索搜尋引擎。將自然語言處理與社群網路分析結合，主要針對微網誌（microblog）上富含簡短（short）、口語（jargon）與表情符號（emoticon）等特性的使用者產生文字內容（User-Generated Content）進行分析。當中幾個重要與有趣的課題包含關係識別（relationship identification）：讓電腦從各種社交文本中（如新聞、社群網站、微網誌與維基百科）自動找出這人與人間的關係（如合作夥伴、競爭對手、情侶關係與師生關係），最著名系統為微軟亞洲研究院所開發的人力方系統；情緒偵測（sentiment detection）則是針對特定使用者或特定關鍵字主題，從微網誌文章以及與朋友訊息互動之資料，自動且即時地（real-time）偵測對應的情緒，常被應用於關鍵字廣告、商品推薦等等；資訊摘要（summarization）的技術則是用來將社群網站的訊息繁複的訊息串中產生合於使用者目的之簡短摘要。

當今社群網路分析的重要課題

在大數據技術的進展，以及資料科學相關技術的逐漸成熟，目前社群網路分析之研究主要由資料探勘領域（Data Mining）所主導，由全球計算機協會所舉辦的頂級知識發現與資料探勘國際學術研討會（ACM SIGKDD），今年 2014 更以 Data Mining for Social Good 為大會主題徵求社群分析相關之創新研究議題與技術突破。以下根據近幾年資料探勘領域之相關論文發表，簡介當今數個社群網路分析最熱門也最重要的研究課題。

1. 資訊散播（Information Diffusion）

Facebook 與 Twitter 與人們生活變得密不可分，其中一個關鍵是這些線上社群服務已成為人們取得資訊與散播資訊的主要管道。社群網站如 Facebook 最根本也最重要的功能是允許使用者對於感興趣的資訊（包含文章、圖片或影片等）進行「按讚」表示喜好、「分享」傳播資訊，與「評論」發表意見，透過這三種方式，加上背後連接數以億計人們的社群網路，資訊散播變成非常容易。我們可以將在社群網路中的資訊散播視為一種類似口碑行銷的影響力傳播（Influence Propagation）過程，一開始資訊會由少數人從外部管道（如新聞媒體或其他社群平台）獲得，這些人稱為種子節點（seeds），接著透過人際關係被傳播，使得更多人獲得資訊（或稱被影響），這些直接或間接從種子節點獲得資訊的人稱為傳遞節點（translators）。病毒式行銷（Viral Marketing）是資訊散播最重要的應用，意即在做廣告

行銷時，由於預算有限，通常只能選擇少數使用者進行行銷，該如何自動地從社群網路中找出少數最具影響力的種子節點，使得資訊以他們為起點開始擴散，能夠達到最多人被影響，這在研究上稱作影響力最大化問題（Influence Maximization）；另一方面，我們可以利用資訊散播來設計廣告宣傳文案：該挑選哪些熱門關鍵字從種子節點擴散，使得影響力能夠最大化，本質上這屬於關鍵字熱門程度預測（popularity prediction）的問題。此外，當資訊擴散時，我們也好奇該主題是否會影響到某些我們所感興趣的人或族群，此問題則稱為散播預測（diffusion prediction）的問題。

2. 連結預測（Link Prediction）

線上社群網路是會隨著時間演變成長的，以 Facebook 為例，在 2011 年有 8 億左右的使用者，到 2013 年成長為 11 億以上的使用者，而新的使用者成員會產生更多的新連結，同時既有成員之間也會形成新的連結，造成社群網路的演化。連結預測之目的在於基於已知的社群網路結構，預測舊成員彼此間如何建立新連結，新成員如何與舊成員建立連結，以了解社群網路演化背後的機制；從應用面的角度，連結預測可以視為一種朋友推薦，即線上社群服務該如何設計準確的預測模型，推薦某些潛在朋友給使用者，使得他們會建立新連結。目前已被驗證為有效的連結預測模型須包含以下幾種特徵：(1) 共同朋友越多越好；(2) 共同興趣越多越好；(3) 居住地之地理距離越近越好；(4) 社群網路上的距離越近越好；(5) 共同朋友的平均朋友數越少越好；(6) 節點越重要越好（如名人）。另一方面，連結預測經常應用於商品推薦，其原理是把建構二元圖（bipartite graph），其中一邊的節點是使用者，另一邊是各種要被推薦的項目（如圖片、音樂與電影），若使用者喜好特定項目，就在使用者與該項目之間建立連結，推薦新項目給使用者就等同基於已知某些使用者喜好，來預測使用者與其他項目之間的連結。

3. 社群取樣與摘要（Sampling and Summarization）

使用者在社群服務上互動所快速累積產生的資料量已達億萬規模，此般巨量資料在資訊擷取、處理與分析等各方面造成嚴厲的挑戰。由於資料量指數成長與快速變動等特性，巨量社交資料通常無法一次完整取得，也因資料規模是 Giga 甚至 Tera，無法讀入記憶體處理，因此如何設計系統性、有效且有效率的方法以取得具有代表性的資料來做後續探勘處理變得十分重要。為儲存與分析超大資料量的社群網路，其中一種方法為簡化（simplification），目的在於使得簡化過後、較為精簡的社群網路仍具有代表性，意即保有原社群網路之各種統計特性，或簡化過程中所產生的誤差越小越好。目前有兩種主流的社群網路簡化技術：取樣與摘要。「取樣」假定完整社群網路無法看見或取得全貌（例如在爬取 Twitter 社群網路時），其目的在於從現有資料中取樣出一個保有特定網路特性（如上述之低平均路徑距離、高群聚係數、與節點度冪次分布）之子圖。「摘要」則假定完整圖形可事先得知，但網路結構太大以致機器無法儲存或肉眼難以辨識，其目的在於從結構或語意上濃縮原圖，使得簡化的圖與原圖之誤差保證落在某一容許範圍之內。

4. 網路社群偵測（Network Community Detection）

在社群網路中，網路社群代表著一群群彼此互動關係密切且有著共同興趣（如）的人，其所代表的意義依據不同類型的社群網路而有所不同，例如在以朋友為基礎的社群網路，網路社群代表的可能是不同興趣的朋友，或不同時期認識的朋友；在公司內部合作關係為基礎之社群網路，網路社群可能是一個個部門；在學術研究共同作者為基礎之社群網路，網路社群可能是不同研究領域。從社群網路圖形結構來看，一個網路社群是一個成員間彼此緊密連接（densely connected）且與其他社群之成員連

結較為稀疏（sparsely connected）的子圖（subgraph），而網路社群偵測之目的是自動地找出存在於一個社群網路的所有網路社群，而最常見的偵測方法，是將問題轉為該如何切割（partition）一個社群網路，使得被切出來的每一個子圖正好是一個網路社群，通常我們會用一個叫模組性（modularity）的衡量指標，來判定一種切割方式好或不好，一個好的偵測方法會得到高的模組性，意即使得所切割出來的社群內部節點彼此有越高密度的連結，且不同社群之間的連結則密度越低。成功地偵測網路社群將可以幫助我們更了解不同社群網路的生態，如不同族群的勢力消長、不同族群的意見領袖、以及哪些人扮演不同族群溝通的橋樑。

結語

雖然上述社群網路分析之相關介紹主要以人與人之間的各種關係與互動為主，但由於社群網路本質是一種描述關聯性資料的圖形結構，因此亦可用來表示其他類型的資訊網路，例如由網頁與超連結所組成的全球資訊網、由論文與引用關係所組成的學術網路、由神經元與神經傳導物質所組成的大腦神經網路等，於是在研究社群網路分析各式問題的同時，我們可以思考該問題對應到其他類型網路的意義是什麼，例如連結預測用於學術網路是文章引用推薦，也可以從其他類型網路來思考社群網路分析潛在的新問題，例如在全球資訊網中偵測情色網站是屬於社群網路的異常偵測（Outlier Detection）或身分解析（Entity Resolution），因而延伸社群網路分析可能的應用。另一方面，因應即接到來的物聯網（Internet of Things）時代，即人們與萬物相互連結形成一個巨大而複雜的異質性網路，此時社群網路分析扮演何種的角色，又該如何與物聯網整合應用，將是社群網路分析下一個十年的趨勢。