知識天地

社群網路分析與應用

楊得年副研究員（資訊科學研究所）

摘要

線上社群網路（Online Social Networks）的盛行，改變了人與人之間互動與資訊傳播的方式。過去相關研究中，多是利用現有社群資料（如DBLP、Twitter）進行分析。常見的做法是將社群網路轉換為圖形，然後分析圖中的特性或是偵測社群網路內值得關心的特殊結構。由於社群網路資料庫已十分完善，且社群應用在生活中辦演角色日益重要，故我們在本文中概述三個嶄新的社群網路應用問題，包含（1）地理資料庫之社群群組查詢處理、（2）多網路拓樸聯合取樣與（3）指定目標使用者的社群影響力最大化。相較於以往社群網路分析的研究，本文中的三個問題針對多種商業與生活上的應用從不同的角度切入，並進行理論推導、演算法設計與實做，以發展具理論深度且實務價值之關鍵技術。

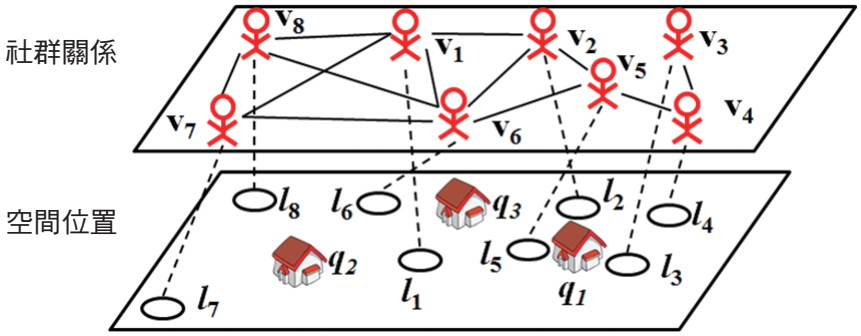
背景介紹

線上社群網路（Online Social Networks）的盛行，改變了人與人之間互動與資訊傳播的方式。根據2011年IEEE Spectrum期刊的報告，線上社群網路（以下簡稱社群網路）獲選為2001年至2010年十一個重大科技的第二位，顯示社群網路對人們的巨大影響。社群網路是由一群行動者（Actor，通常是個人或是組織）及行動者之間的關係（Relationship）所構成。早期的社群網路研究常見於社會科學領域中，然隨著社群網路的規模日漸增加，且裡面包含的元素日益複雜（例如原本每個節點只是一個使用者，但現在可能還包含了許多關於此使用者的描述，或是該使用者產生和分享的多媒體資料），故需要仰賴電腦科學的方式，對豐富多元的社群網路進行有效率地分析和探勘。社群網路於電腦科學領域相關研究中，社群網路分析（Social Network Analysis）是一個非常熱門的議題。過去相關研究中，多是利用現有社群資料（如DBLP、Twitter）進行分析。常見的做法是將社群網路轉換為圖形，然後分析圖中的特性。社群網路分析常用基本指標有包括密性（Closeness）、中介性（Betweenness）及中心性（Centrality），以及一些延伸的社群網路特質，如小世界（Small World）和冪律邊度數分佈（Power-Law Edge Degree Distribution）。另外，也有許多研究著重於偵測社群網路內值得關心的特殊結構，例如找出社群網路中團體的變化。傳統社群研究多探討本質性問題，而由於社群網路資料庫已十分完善，且社群應用在生活中辦演角色日益重要，本文將介紹三個嶄新的社群網路應用問題，包含（1）地理資料庫之社群群組查詢處理、（2） 多網路拓樸聯合取樣與（3）指定目標使用者的社群影響力最大化。地理資料庫之社群群組查詢結合社群網路的結構與使用者的空間位置，以找出一群適當的人選與聚會地點進行即時性的聚會活動，或是提供商家進行適地性的行銷服務；多網路拓樸聯合取樣同時取樣多個社群網路，透過現有之網路帳號匹配演算法進行整合，以提供統計上保證所取樣與整合之多網路拓樸與真實網路拓樸差異。藉由整合多網路拓樸與統計上之保證，取樣出來的網路拓樸可增進社群網路分析演算法的準確性。最後，指定目標使用者的社群影響力最大化問題討論如何透過選擇適當的中介使用者，讓某個起始使用者對某個目標使用者的影響力最大化，達到行銷的目的。

地理資料庫之社群群組查詢處理

現有之地理資料庫查詢處理多假設每一物件為一地點，如加油站、遊樂區，或者是可移動之車輛，並未考慮每一物件可能為一人員。然而，隨著社群網路之興起，目前許多社群服務網站，如Facebook Places、Meetup、Geomium、Buddy Beacon與FindMe等，均允許使用者隨時透過行動裝置上傳當下GPS座標，以分享給好友。然而，使用者雖然隨時知道好友的位置，但在即時安排社群活動時，仍得自行決定參與者。對於即時性活動，使用者距離近之好友往往擁有較小之交通時間，故較為適合。然而，選擇上述好友並不代表上述參與者之間均熟識，另一方面，選擇熟識的參與者並不代表他們離聚會地點的空間距離接近。因此，如何選擇適當的參與者，與適當的聚會地點，讓參與者彼此之間熟識且他們距離聚會地點的距離（交通時間）很小以利進行聚會是一個很有挑戰性的問題。

此外，目前團購網站如Groupon已可根據使用者過去的團購紀錄來推薦適當的團購項目（Groupon Personalized Deal）。然而，該功能只能推薦使用者可能感興趣的項目，而不能推薦使用者可以和哪些朋友一起參與該項團購。因此，藉由我們提出的查詢處理系統，團購網站不僅能推薦產品，同時亦能推薦可以一起團購的朋友，如此不僅能提高團購的成功率，並能增進朋友間之互動與社交關係。例如目前使用者若希望透過團購網站來購買某餐廳之折價卷，目前上述團購網站的推薦功能僅能推薦其可能感興趣的餐廳折價卷給使用者。在整合本查詢處理後，可推薦使用者與一起團購的好友距離他們不遠的餐廳折價卷，讓他們能一起用優惠的價格在鄰近的餐廳進行聚會。

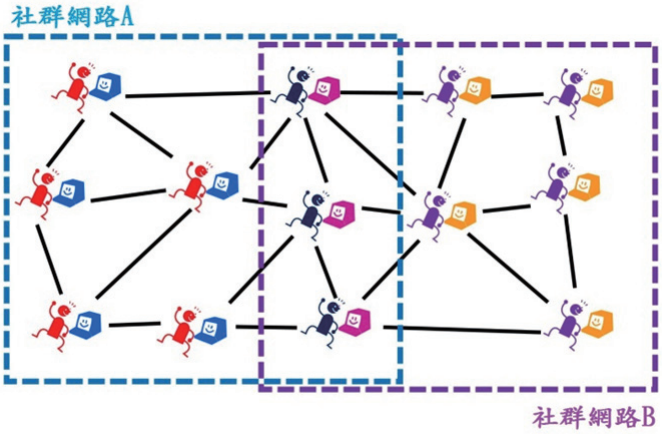


圖一、社群關係與空間位置

圖一為用以說明同時考量空間距離與社群關係的挑戰性與複雜度的範例。圖一為八位備選成員其社群關係圖與空間位置。於社群關係中，人與人之間的線條表示兩者熟識（如v1與v6），而每個人的空間位置亦顯示於圖一中（l1~l8）。圖中亦有三個備選的聚會地點（q1,q2,q3）。假設我們要找四個參與者，且限制這四個參與者每個人至多不認識其他一位參與者，同時亦要找出距離這四個參與者最近的聚會地點。一種可能的方法是只根據空間距離來選擇，舉例來說，{v2,v3,v4,v5}距離聚會地點q1最近。然而，這四個參與者並不符合上述之社群熟悉度條件（每個人至多能不認識另外一個參與者），這可能會讓聚會的氣氛變差。另一方面，如果只看社群關係，則可能會選出{v1,v6,v7,v8}與聚會地點q2。然而，雖然這四個人彼此間皆熟識，但他們距離聚會地點q2的空間距離很大，故不適合即時性的聚會。相較之下，{v1,v2,v5,v6}配合聚會地點q3是一個較佳的選擇，因為這四位參與者每位至多不認識另外一位參與者，同時他們至聚會地點q3的空間距離亦很小。

為了探討上述結合空間距離與社群關係的即時性聚會問題，我們提出了以下的問題：給予多個備選聚會地點之地理位置、每一位使用者當下地理位置，以及使用者之間社群關係，欲從社群網路中求得大小為p之社群群組，我們同時針對群組成員與集合地點進行考量，使得被選出的這p個群組成員中且每一成員最多僅能允許不認識k位其他成員，同時，亦從備選聚會地點中選出一個聚會地點，使得該群組成員至聚會地點之地理距離總合最小。有效率地找出這p個群組成員與聚會地點十分困難，因為要同時考量群組成員人數、是否認識與空間距離。我們證明該問題是NP-Hard且不存在任何近似演算法，並提出一個有效率的搜尋演算法，藉由空間距離和群組成員間的社群關係，配合R-Tree與BallTree空間索引結構，推導出多種不同的限制函數，藉此來針對引導搜尋樹優先展開較好的可行解，並配合這些可行解來修剪搜尋樹，以大幅減少搜尋演算法所需的時間，並可在很短的時間內取得最佳解。我們亦將該查詢系統實做於Facebook上，並進行使用者研究。結果顯示，我們的演算法不僅可很有效率地找到最佳解，同時使用者亦同意本查詢系統所找出的群組成員較使用者自行選出的群體成員更適合即時性的聚會。

多網路拓樸聯合取樣



圖二、社群網路交集

社群網路分析為研究熱門主題，現有之社群網路分析演算法與技術大部分皆於社群網路取樣出之真實資料集上面進行評估與測試，而常用之取樣演算法包含廣度優先搜尋取樣、隨機漫步取樣與均勻取樣。然而，目前現有之取樣演算法皆未提供統計分析來保證所取樣之網路拓樸與實際差異。是故，對於研究者而言，十分難去評估其所取樣網路拓樸品質好壞。更重要的是，由於目前社群網路各自獨立，如圖二所示，社群網路A可為Facebook，社

群網路B可為LinkedIn，中間交集為Facebook與LinkedIn皆有帳戶之使用者。相同使用者於不同社群網站所連結到之鄰近點（即朋友）並不相同，故目前線上社群網路並非一完整之網路拓樸；換言之，每個社群網站拓樸所紀錄的相鄰點，均僅為使用者真實世界之一小部分朋友。是故，目前所有社群網路分析之相關文獻對Facebook、DBLP、Twitter等真實社群拓樸所進行之實驗，受限於上述網路均僅為完整線上社群網路之子圖，故探討所得之分析結論並不代表真實社群網路之性質。幸運的是，隨著社群網路分析演進，近年來出現許多研究提供如何利用使用者帳號與其網路拓樸特徵將不同社群網路下，相同之使用者做匹配，像是個別使用者由於所用單字由於人腦限制，所創帳號彼此之間關聯性很強。或是利用網路社群特徵，舉凡像是共同朋友、杰卡德係數等等。

因此，我們提出一個新的系統框架以同時取樣多個社群網路，透過現有之網路帳號匹配演算法進行整合。我們主要目標在於能夠提供統計上保證所取樣與整合之多網路拓樸與真實網路拓樸差異小於某個閥值，打破過去網路拓樸採樣所公開之資料中皆未提供其採樣品質保證之缺陷。而這裡所謂真實完整網路拓樸代表的是在所有社群網路拓樸中，將相同使用者帳號完全匹配對應起來，所整合出完整社群網路。單一網路下，如果（1）取樣節點越少且（2）取樣演算法非均勻，其取樣之網路拓樸特性會遠離完整網路拓樸。然而多網路拓樸底下並不見得成立，若某網路拓樸取樣過多時，最後整合之多網路拓樸會趨近於此網路拓樸，也會造成偏差，使其取樣之網路拓樸特性會遠離完整網路拓樸。而一種簡單取樣作法是根據各網路拓樸大小取樣，假設社群拓樸A共有10000個使用者，社群拓樸B共有20000個使用者。我們即依照1:2之比例，對兩個網路拓樸做取樣，然而這樣簡單的取樣方法並非最有效率之方法，取樣品質相關因素不僅是原本網路拓樸大小，帳號匹配演算法之正確性亦影響取樣與整合之多網路拓樸與真實網路拓樸差異，而不同社群網路帳號對應之困難度亦不同，如何依照現有之帳號匹配演算法正確率，進一步調整各社群網路取樣點數使取樣之網路拓樸特性不會遠離完整網路拓樸亦為一大難題。相較於單一網路取樣，如何在多網路拓樸架構下清楚分析取樣結果更加具有挑戰性與應用價值。

由於多網路品質分析相當複雜，需要系統性分析，因此，我們按部就班一一加入各個因素進行分析，達成多社群網路拓樸聯合取樣之數學模型建置，從雙網路拓樸開始，按部就班分析（1）無交集（2）交集且有完美匹配演算法與（3）交集且非完美匹配演算法。再將此模型推廣至多個網路拓樸取樣。而針對取樣策略：（1）規模限制取樣（2）品質限制取樣（3）時間限制取樣。我們根據上述分析結果，對於三種取樣策略進行最佳化，使理論之取樣結果與真實完整網路拓樸之差異最小化。

指定目標使用者的社群影響力最大化

社群影響力是近年相當重要的研究議題。2012年的Nature期刊中的實驗顯示，知名社群網站Facebook上的使用者，確實會因其Facebook好友的社群影響力而改變自身行為。如何利用社群影響力來影響使用者購買的行為，達到行銷目的，這種新型態的社群行銷方式，或許是近年最有效的廣告方式。隨著社群網路服務如Facebook、Twitter等的發展，社群行銷的重要性與日俱增，知名媒體USA Today、BBC News、Bloomberg Businessweek、New York Times皆曾撰文指出社群行銷的重要性。然而，過去文獻針對口耳相傳（Word-of-Mouth）多著重於選擇具影響力的使用者以對擴散範圍最大化，對於單一特定的目標使用者，如何使其受到之社群影響力總合最大化並無著墨。

中介使用者（Intermediate User）的選擇對社群影響力的傳遞來說相當重要。因此，我們探討一個新的社群影響力最大化問題：APM（Acceptance Probability Maximization），給定一位起始使用者、一位目標使用者、與其中介使用者總數上限，APM 會選擇適當的中介使用者集合，使起始使用者對目標使用者的社群影響力最大化。我們證明於獨立串聯（Independent Cascade）模型中，計算社群影響力大小是 #P-hard 的問題，且在獨立串聯模型中的APM是NP-hard 問題。因此，我們利用另一最大影響力子樹（Maximum Influence Arborescence）模型來模擬社群影響力傳遞。在此模型下，我們提出多項式演算法演算法，利用動態規劃在多項式時間內獲得 APM 的最佳解。

我們亦探討在存在時間限制的情況下，對於給定起始使用者、目標使用者及中介使用者總數上限，如何選擇適當的中介使用者集合，使得起始使用者對目標使用者在時限內的影響力最大化。值得注意的是，在考慮時間因素的情況下，使用者在社群網站上的活躍程度與在線時間，也與其否能在時限前影響其他人的決定有重大關連。舉例來說，通常下班後至睡前在線的機率會較大，而睡眠時間的在線機率則較小，另外，由於使用者所在時區的不同，在某特定時間點在線的機率也會不同。此問題的難點主要在於中介使用者的選擇。一般說來，選擇社群影響力較大的中介使用者會增加對社群影響目標對象的影響力，但若影響力較大的中介使用者較不活躍，那麼選擇一個社群影響力（或同質影響力）較小但活躍的使用者，會讓時限內社群影響目標接收到此訊息而接受的機率增加。此外，另一個難點是決定起始使用者發送訊息給各中介使用者的時間，若訊息太早發送給一個中介使用者，若其鄰點尚未收到訊息，便無法借重其鄰點的社群影響力使該使用者接受；但若等待該中介使用者太久，則剩餘時間可能不足以得到社群影響目標的回覆。因此，社群影響力及時間兩個維度需要同時被考慮，增加了此問題的計算難度。我們提出多項式演算法，利用動態規劃在多項式時間內獲得此問題的最佳解。