前言

以下的事情都有共通點：市場歸因分析、反洗錢（AML）分析、客戶旅程建模、安全事件因果因素分析、文獻探勘、詐欺網路檢測、網際網路搜索節點分析、地圖應用建立、疾病聚類分析，以及分析一齣 William Shakespeare（威廉．莎士比亞）的戲劇。正如你可能已經猜到的，所有這些事情的共同點是圖形，佐證了莎士比亞「世界就是一張圖形（All the world's a graph!）」的說法。

好吧！莎士比亞那句話中寫的不是圖形，他寫的是舞臺（譯按：世界就是一座舞臺）。不過，請注意，上面列出的事情都包含實體及實體之間的關係，關係又包括直接和間接（傳遞）關係。實體是圖中的節點——它們可以是人、事件、物件、概念或位置。節點之間的關係則是圖形中的邊（edge）。因此，莎士比亞戲劇的本質不就是鮮明描繪實體（節點）及其關係（邊）嗎？ 因此，也許莎士比亞可以用「圖形」二字來改寫他那著名的宣言。

讓圖形演算法和圖形資料庫如此有趣和強大的，不是兩個實體之間的簡單關係（A 與 B 之間存在關係）。畢竟，資料庫的標準關聯式模型的基礎，實體關聯圖（entity relationship diagram，ERD）中，早在幾十年前就實例化了這種類型的關係。使圖形如此重要的是方向關係和傳遞關係。在方向關係中，A 可能導致 B，但反過來不行。在傳遞關係中，A 可以與 B 直接相關，B 可以與 C 直接相關，但 A 和 C 沒有直接關係，所以 A 和 C 有傳遞關係。

碁峰 www.gotop.com.tw

有了傳遞關係後——特別是當這些關係數量很多而且種類多樣，具有許多種可能的關係 / 網路樣式，實體間的不同的分支展開程度——圖形模型可以揭示關聯式資料庫無法偵測到的實體間關係，即使實體間看起來沒有連接或是毫不相關。因此，圖形模型可以有效地應用於許多網路分析情境中。

參考一下這個行銷歸因的情境用例：A 看到行銷活動；A 在社交媒體上談論它；B 與 A 有聯繫並看到評論；然後，B 購買產品。從市場行銷活動經理的角度來看，標準的關係模型無法確定歸因，因為 B 沒有看到該活動，而 A 沒有對該活動做出回應。活動看起來像失敗，但它的實際成功（和正面的投資回報率）是透過行銷活動和最終客戶購買之間的傳遞關係，透過中間人（中間的實體），由圖形分析演算法發現的。

接下來，參考一個反洗錢（AML）分析案例：A 和 C 涉嫌非法販毒。這兩人之間的任何互動（例如，金融資料庫中的金融交易）都將被當局標記，並受到嚴格審查。但是，如果 A 和 C 從未一起交易業務，而是透過一位安全、無不法記錄和沒有被金融標記的 B 進行金融交易，那麼該如何挑出這樣的交易呢？用圖形分析演算法！圖形引擎會透過中間人 B 發現 A 和 C 之間的傳遞關係。

在網際網路的搜索中，主要的搜尋引擎使用超連結網路（以圖形為基礎）。演算法在整個網際網路上查找任何給定搜索詞的中心權威節點。在這種情況下，邊的方向性至關重要，因為網路中的權威節點也是許多其他節點指向的節點。

文獻探勘（LBD）——是一個知識網路（以圖形為基礎）的應用，它能夠在數千（甚至數百萬）篇研究期刊文章的知識庫中找到重大的發現——「隱藏的知識」只有透過已發表的研究結果之間的關係才能被發現，這些研究結果可能有很大的分支展開（傳遞關係）。LBD 正被應用於癌症研究，在癌症研究中，症狀、診斷、治療、藥物相互作用、遺傳標記、短期結果和長期後果中所含大量語義醫學知識，可能「隱藏」以前未知的治療方法或對難症病例的有益治療。知識可能已經存在於網路中，但我們需要把線索串連起來才能找到它。

前面列出的情境用例中，所有透過圖形演算法進行網路分析的範例，都可以看出圖形化的威力。每個案例都與實體（人、物件、事件、行動、概念和地點）及其關係（接觸點，因果關係和簡單關係）密不可分。

在考慮圖形化的威力時，我們應該記住，在現實世界的圖形模型中，最強大的節點可能是「上下文（context）」。上下文可能包括時間、位置、相關事件、附近的實體等等。將上下文合併到圖形中（作為節點和邊），可以製造出令人印象深刻的預測分析和診斷性分析能力。

碁峰 www.gotop.com.tw

Mark Needham 和 Amy Hodler 的圖形演算法這本書，旨在拓寬我們對這些重要類型的圖形分析的知識和能力，包括演算法、概念和落實在機器學習應用。從基本概念到重要的演算法，再到處理平台和實際用例，作者為美妙圖形世界編寫了一本兼且指導性和說明性的指南。

— *Kirk Borne, PhD*

Principal Data Scientist 和 Executive Advisor

Booz Allen Hamilton

March 2019

碁峰 www.gotop.com.tw

第一章

引言

圖形是電腦科學的共通主題之一，它是一種描述運輸系統、人與人之間互動和電信網路組織的抽象表示。許多不同的結構，都可以使用同一個形式來建模，提供學過圖形的程式師一個巨大的力量。

—演算法設計手冊作者 *Steven S. Skiena*（*Springer*）

石溪大學（*Stony Brook University*）資訊工程傑出教學教授

現今在資料處理上，最大的挑戰都集中在關係處理，而不只是將離散資料做成表格而已。學術研究、社會活動和商業行為使用了連結性資料，圖形技術和圖形分析為這類連結性資料提供了強大的工具，例如：

- 為金融市場到資訊科技的變動環境進行建模
- 預測傳染病的傳播，以及服務延遲與中斷
- 為機器學習找尋預測的特徵值，以打擊金融犯罪
- 找出個人化體驗與推薦

由於現今資料間的連結性和系統的複雜度提升，所以使用資料中豐富且不斷發展的關係，是很重要的課題。

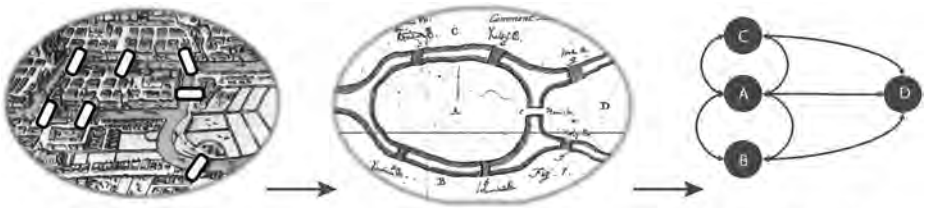
本章介紹了圖形分析和圖形演算法。在介紹圖形演算法和解釋圖形之間的區別以及圖形處理之前，我們將簡短地介紹圖形的起源。我們將探討現代資料本身的性質，以及資料連接關係中包含的資訊如何比我們用基本統計方法能發現的更為複雜。本章也會歸納哪些地方適用圖形演算法。

碁峰 www.gotop.com.tw

什麼是圖形？

圖形的歷史可以追溯到 1736 年，出現在當李昂哈德．歐拉（Leonhard Euler）解決了柯尼斯堡（Königsberg）七座橋的問題時。該問題是，柯尼斯堡有四個由七座橋相連的區域，請問是否有可能藉由七座橋走遍所有區域，而且每座橋只能走過一次，這個問題的答案是不可能。

歐拉認為問題只和連接相關，因此為圖論及其數學奠定了基礎。圖 1-1 用一張歐拉的原始草圖描述他的做法，該草圖來自於論文 *Solutio problematis ad geometriam situs pertinentis*（*http://bit.ly/2TV6sgx*）。



走遍柯尼斯堡的橋
柯尼斯堡有四個由 7 座橋連接的主要區域。請問你可以走遍所有的橋，並回到你的出發地嗎？

歐拉的看法
所有相關的資料，只有區域和其間相互連通的橋而已。

圖論的起源
歐拉將問題抽象化，並基於點和關係建立起泛用的通則，這樣的通則可以適用任何有連結關係的系統。

圖 1-1 圖論的起源。柯尼斯堡（*https://bit.ly/2JCyLvB*）城市中，包括兩個相互連接的大島和兩個由七座橋連接的的主要陸地。謎題是創造一個步行穿過城市的路徑，這個路徑要跨越每座橋一次，而且只能一次。

圖形雖起源於數學，也是一種實用的、高保真的資料建模和分析方法。構成圖形的物件稱為節點或頂點，它們之間的連結稱為關係、連結或邊。我們在這本書中使用了**節點**（*node*）和**關係**（*relationship*）這兩個術語：你可以把節點看作句子中的名詞，把關係看作節點與節點間的動詞。為了避免混淆，我們在這本書中討論的圖與圖 1-2 中的那種圖形方程式或圖表沒有任何關係。

請看圖 1-2 中的人物圖，我們可以很容易地造出幾個句子來描述它。例如，A 和擁有一輛車的 B 住在一起，而 A 駕駛的是 B 的車。這種建模方法很有吸引力，因為它很容易映射到現實世界，並且非常容易畫出來，這一點有助於保持資料建模和分析間的一致性。

碁峰 www.gotop.com.tw

但是做出建模圖只是故事的一半。我們可能也希望透過一些處理，來揭示出不明顯的資訊，這便是圖形演算法的領域了。

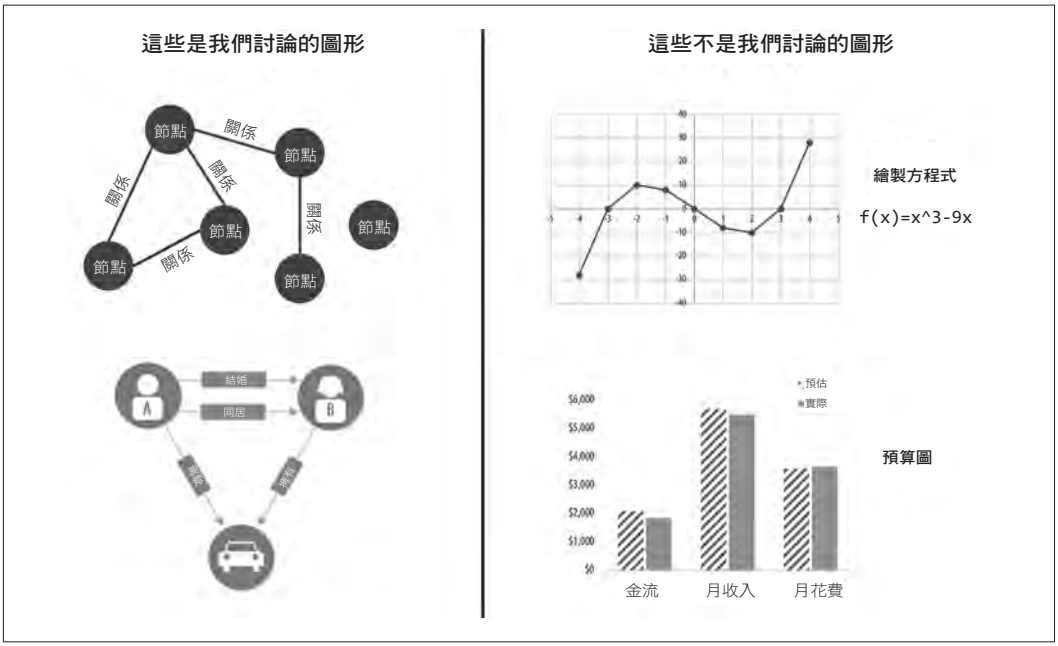


圖 1-2 圖是一種網路的表示法，通常用圓圈表示節點，用線條表示關係。

什麼是圖形分析和圖形演算法？

圖形演算法是圖形分析工具的子集。圖形分析是一種動作，它使用任何基於圖形的方法來分析連接的資料。我們可以使用各種方法：查詢圖形資料、使用基本統計資料、視覺化地瀏覽圖表，或者將圖形合併到我們的機器學習任務中。圖形樣式查詢通常用於本地資料分析，而圖形計算演算法通常涉及更多的全域和迭代運算分析。這些分析類型在用途上儘管存在著重疊，但我們使用術語**圖形演算法**（*graph algorithm*）來代表後者，即含更多的計算分析和資料科學的用途。

碁峰 www.gotop.com.tw

網路科學

網路科學是一個根植於圖論、涉及物件間關係數學模型的學術領域。由於資料的大小、連通性和複雜性，網路科學家依賴於圖形演算法和資料庫管理系統。對於複雜性和網路科學，有許多很棒的資源，這裡有一些參考資料供您探索。

- *Network Science*（*http://networksciencebook.com/*）由 Albert-László Barabási 編寫，是一本入門型電子書
- Complexity Explorer（*https://www.completityexplorer.org/*）提供線上課程
- The New England Complex Systems Institute（*http://necsi.edu/*）提供各種資源和論文

圖形演算法是分析連結資料最有效的方法之一，因為它的數學計算是專門為處理關係而建構的。圖形演算法描述了一些步驟，這些步驟用來處理圖形，以發現其一般性質或特定數量的步驟。基於圖論的數學基礎，圖形演算法利用節點之間的關係來推斷複雜系統的組織和動態。網路科學家使用這些演算法來發現隱藏的資訊、測試假設，以及對行為進行預測。

圖形演算法具有各式各樣的潛力，從防止詐欺和優化呼叫路由到預測流感的傳播。例如，我們可能希望為電力系統中的超載節點作評分，或者我們可能想要在圖形中根據擁塞程度，為傳輸系統節點做分組。

事實上，在 2010 年時美國航空旅行系統經歷了兩個嚴重的事件，這兩個事件涉及多個繁忙的機場，後來的研究就是使用了圖形分析。網路科學家 P. Fleurquin、J. J. Ramasco 和 V. M. Eguíluz 使用圖形演算法確認事件是系統級延遲的一部分，並將此資訊用於糾正建議，這些資訊描述在他們的論文 *Systemic Delay Propagation in the US Airport Network*（*https://www.nature.com/articles/srep1159/*）中。

為了使空中交通的網路形象化，Martin Grandjean 在他的文章 *Connected World: Untangling the Air Traffic Network*（*http://bit.ly/2CDdDiR*）中創建了圖 1-3。這幅圖清楚地展示了航空運輸集群高度連接的結構。許多交通系統都呈現出連結集中分佈的情況，這種連結情況有明顯的軸輻模式（譯按：星狀網路拓樸），這種模式將影響延遲的狀況。

碁峰 www.gotop.com.tw

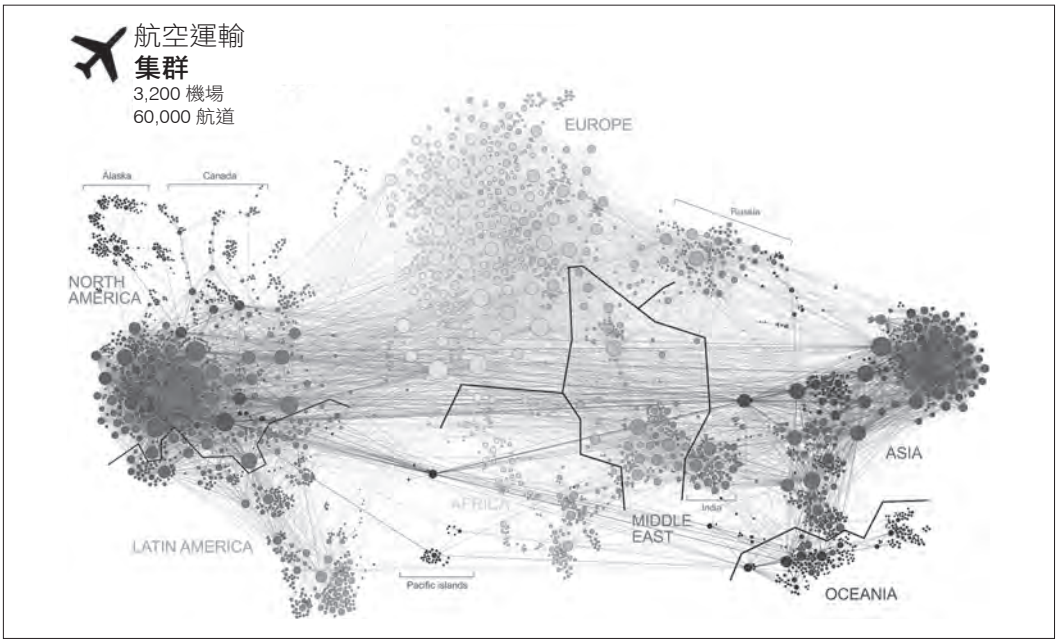


圖 1-3 航空運輸網路展示了多種大小的軸輻結構，這些結構決定了運輸流。

圖形也有助於揭示極小的互動和變動如何導致全域突變。它們將微觀和宏觀尺度聯繫在一起，精確地呈現出全域結構中哪些東西有互動關係。這些關聯用於預測行為和找出缺少的連結。圖 1-4 是一個草原物種間的食物鏈，它使用圖表分析來評估階層組織和物種間的相互作用，並且預測缺失的關係，在 A. Clauset、C. Moore 和 M. E. J. Newman 的論文 *Hierarchical Structure and the Prediction of Missing Links in Network*（*https://www.nature.com/articles/nature06830*）中有詳細描述。

碁峰 www.gotop.com.tw

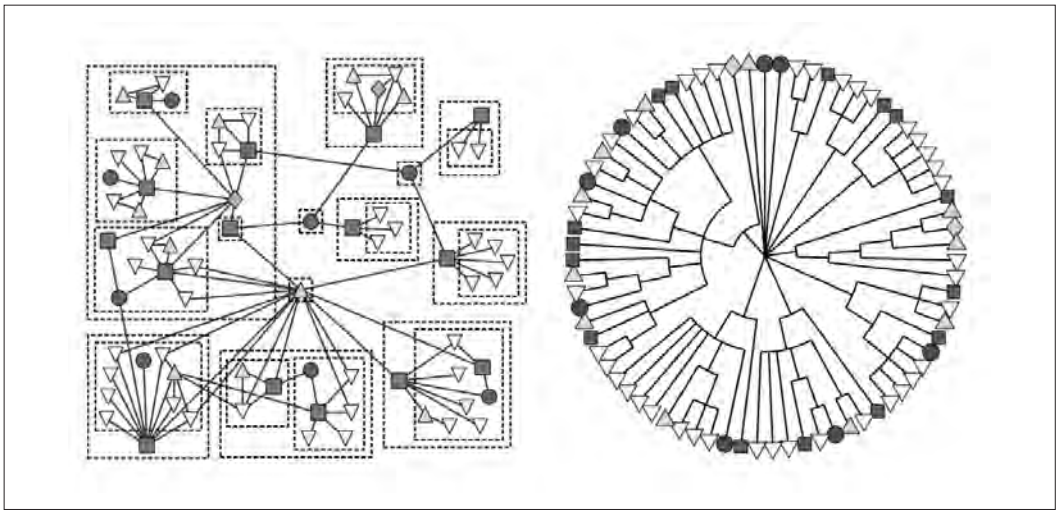


圖 1-4 草原物種的食物鏈圖，使用圖形將小規模的相互關係與大的結構連結起來。

圖形處理、資料庫、查詢和演算法

圖形處理有多種方法，以圖形工作流程與工作種類區分。多數的圖形查詢都牽涉了圖的一部分（例如，起始節點），而工作通常集中在周圍的子圖形中。我們將這種類型的工作圖稱為**圖形局部**（*graph local*），它意味著宣告式查詢一個圖的結構，正如 Ian Robinson、Jim Webber 以及 Emil Eifrem 在 *Graph Databases*（O'Reilly）一書中所解釋的那樣。這種局部圖形的處理常用於即時事務和樣式為基礎的查詢中。

在談到圖形演算法時，通常會尋找全域模式和結構。演算法的輸入通常是整個圖，輸出可以是一個加了更多料的圖或一些聚合過的值，如一個評分。我們將這種處理歸類為**圖形全域**（*graph global*），它通常使用計算型的演算法（通常要多次迭代）處理圖形的結構。這種方法透過網路連接揭示了網路的整體性質。組織傾向於使用圖形演算法來建模系統，並根據事物的傳播方式、重要元件、分組標識和系統的整體強健性來預測行為。

這些定義可能有些重疊的部分，也就是說有時我們可以使用某種演算法處理來回應局部查詢，但也可以用同一種演算法處理來回應全域問題。但簡單地說，計算型演算法的處理是對整個圖進行操作，而資料庫中查詢的則是局部圖形。

傳統上，交易處理和分析都是被分開的，這是一種基於技術限制的非自然區隔。我們的觀點是，圖形分析能提升更聰明的交易處理，這為進一步分析創造了新的資料和機會。最近出現了一種趨勢，將交易處理和分析結合起來，達成即時的決策。

碁峰 www.gotop.com.tw

OLTP 和 OLAP

線上交易處理（*online transaction processing*，OLTP）操作通常是短期活動，如預訂機票、查詢帳戶額度、預訂銷售等等。OLTP 意味著大量的低延遲查詢處理和高資料完整性。儘管 OLTP 每個交易可能只涉及少量記錄，但系統可是同時處理許多事務的呢！

線上分析處理（*online analytical processing*，OLAP）有助於對歷史資料進行更複雜的查詢和分析。這些分析可能包括多個資料來源、格式和類型。檢測趨勢、進行假設分析、預測和發現結構模式都是典型的 OLAP 的應用。與 OLTP 相比，OLAP 系統處理的交易次數較少，但每次交易包含許多記錄執行時間較長。OLAP 系統傾向於快速讀取，但不像 OLTP 過程中會做交易更新，OLAP 常會進行批次處理操作。

然而，最近 OLTP 和 OLAP 之間的界線開始變得模糊。現代資料密集型應用程式將即時交易操作與分析相結合。會有這種處理上的合併是由軟體發展造就的，而軟體發展又是由幾件事情所推動的，例如可擴展的交易管理和與日俱增的串流處理，以及成本更低、硬體配備更多記憶體。

將分析和交易結合在一起，使得分析工作成為日常的操作之一。隨著資料從銷售點（POS）機器、製造系統或物聯網（IOT）設備被收集，在處理過程中分析工作就能做出即時推薦和決策。這一趨勢是在幾年前就已被觀察到了，用來描述這種合併的術語包括 *translytics*（譯按：組合字，取交易的字首和分析的字尾）和**混合交易分析處理**（*hybrid transactional and analytical process*，HTAP）。圖 1-5 說明了如何使用唯讀副本將這些不同類型的處理結合在一起。

根據 Gartner 的說法（*https://gtnr.it/2FAKnuX*）：

> [HTAP] 可能會重新定義一些業務流程的執行方式，因為即時進階分析（例如，計畫、預測和假設分析）成為流程本身的一個組成部分， 而不是事後執行的單獨活動。這將使即時業務驅動決策處理流程變成一種新的形式。最終，HTAP 將成為智慧業務運營的關鍵支援架構。

碁峰 www.gotop.com.tw

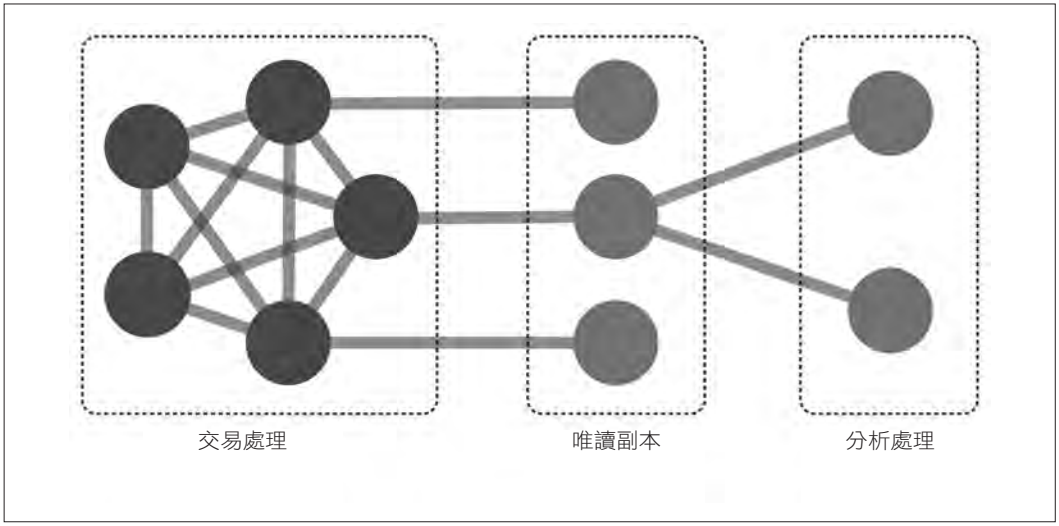


圖 1-5 支援低延遲查詢處理和高資料完整性交易，同時又整合大量複雜資料分析的一個平台。

隨著 OLTP 和 OLAP 間的整合越來越緊密，並開始支援以前只有對方才提供的功能，所以現在不再需要為這些工作使用不同的資料產品或系統；我們可以設計讓兩者使用相同的平台來簡化架構。這代表著我們的分析查詢可以利用即時資料，而且我們將分析反覆運算過程也納入流程中。

為什麼我們要關心圖形演算法？

圖形演算法讓連結性資料更有意義。我們都可以從這些現實系統中看到關係。從蛋白質作用機制到社交網路，從通信系統到電力網路，從零售體驗到火星任務規劃。瞭解網路及其內部連接為洞察和創新提供了難以置信的潛力。

圖形演算法特別適合用來理解高度連接資料庫中的結構，以及揭露其中資料模式。沒有什麼比大數據資料更有明顯的連結性和相互關係了，其彙集、混合和動態更新的資訊量都很驚人。這就是圖形演算法能派上用場的地方，透過利用關係做更複雜的分析，並且增強人工智慧中的資訊量，讓巨量資料變得更有意義。

由於我們的資料連結性變高了，瞭解資料的關係和相互依賴性變得越來越重要。研究網路成長的科學家指出，隨著時間的推移連結性越來越高，但這種成長並不是均勻的成長。偏好依附原則（preferential attachment）是研究動態成長會影響結構的理論之一。如圖 1-6 所示，偏好依附原則指的是當一個節點要加入網路中時，會偏好連結到已經有很多連接的節點。

碁峰 www.gotop.com.tw

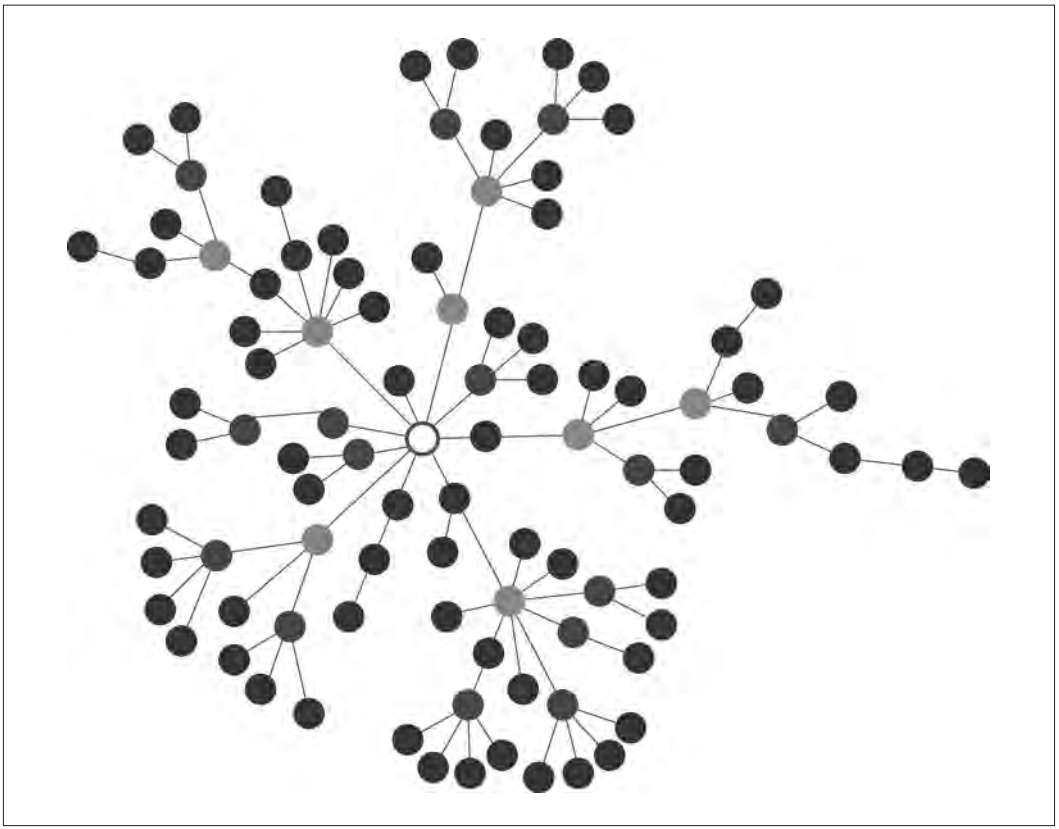


圖 1-6 偏好依附原則是一個節點的連接越多，接收新連結的可能性就越大的現象，會導致集中度上升和中心的不均勻。

在 Steven Strogatz 的書 *Sync: How Order Emerges from Chaos in the Universe, Nature, and Daily Life*（Hachette）中，他提供了例子，並解釋了現實生活系統自我組織的不同方式。不管潛在的原因是什麼，許多研究人員認為，網路的發展與其產生的形狀和層次結構密不可分。高度密集的群體和塊狀資料網路往往會發展起來，同時伴隨著資料規模及複雜性增加。現在我們看到了大多數現實世界網路都是這種關係集群，從網際網路到社交網路，如圖 1-7 所示的遊戲社群。

碁峰 www.gotop.com.tw

圖 1-7 中的網路分析是由 Pulsar 的 Francesco D'Orazio 創建的，目的是預測病毒傳播與傳播策略。D'Orazio 發現（*https://bit.ly/2CCLlVI*）社群分佈集中度與內容傳播速度之間的相關性。

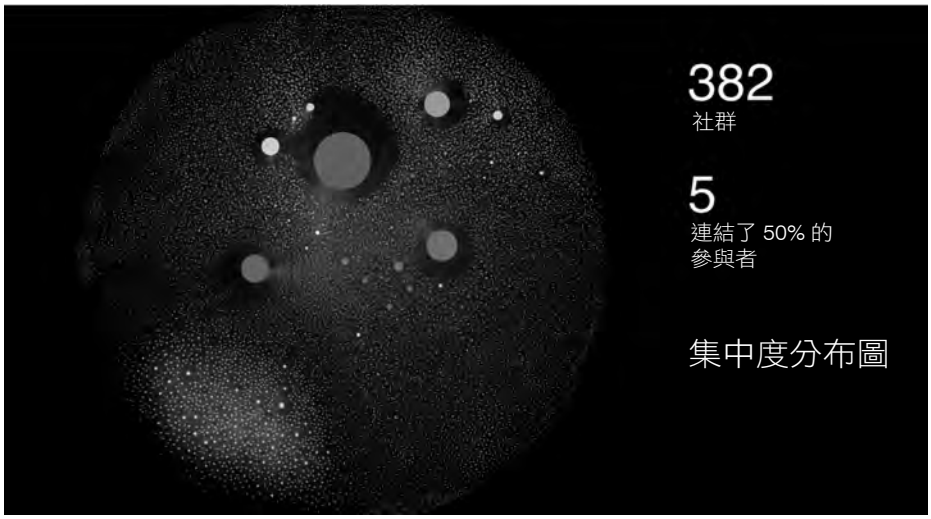


圖 1-7 遊戲社群分析（*https://bit.ly/2CCLlVI*）顯示在 382 個社區中連接只集中在 5 個社群。

這和以常態分佈模型作預測會有很大的差異，因為常態分佈模型下大多數節點的連接數都相同。例如，如果全球資訊網（WWW）的連接呈常態分佈的話，那麼所有頁面進出的連結數將大致相同。常態分佈模型認為，大多數節點是相等連接的，但許多不同種類的圖形和許多實際網路卻表現出集中的情況。不管是網站的圖形、旅遊和社交網路的圖形都一樣，呈現冪次定律分佈（冪律分佈，*power-law* distribution），其中一些節點有著高度連接，但大多數節點只有少數的連接。

冪次定律

冪次定律（也稱為**定標定律**（*scaling law*））描述兩個量之間的關係，其中一個量隨另一個量的冪次而變化。例如，一個立方體的面積與其邊長之間的關係為 3 冪次方。一個眾所周知的例子是**帕累托分佈**（*pareto distribution*）即「80/20 法則」，最初用於描述 20% 的人口控制 80% 的財富的情況，自然世界和網路中可以見到各種冪次定律。

碁峰 www.gotop.com.tw

在觀察關係或是做預測時，試圖「平均」一個網路，不會得到太好的效果。我們很容易從圖 1-8 中看出來，誤用不平衡的資料的平均特徵，將導致不正確的結果。

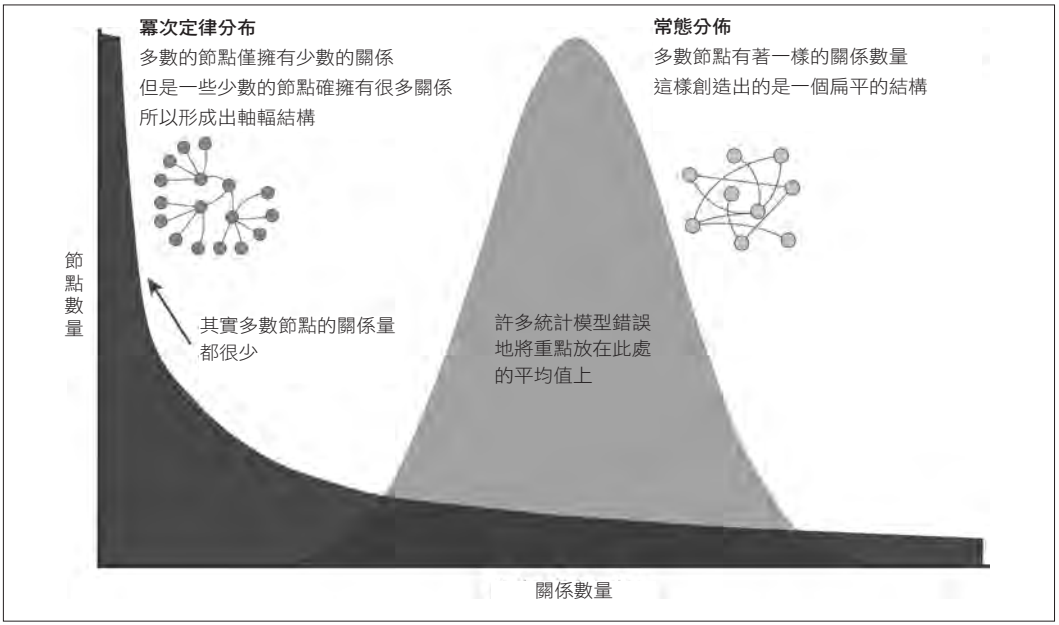


圖 1-8 現實世界中的網路節點和關係分佈不均勻，關係呈現一種極端的冪次定律分佈。常態分佈假設大多數節點在任何網路中具有相同數量的關係。

由於高度連接的資料不遵循常態分佈，網路科學家使用圖形分析去探索和解釋現實世界資料中的結構和關係。

> 我們所知道的自然界中沒有一個網路可以用隨機網路模型來描述。
>
> —東北大學複雜網路研究中心主任、眾多網路科學書籍的作者 *Albert-László Barabási*

多數人會面臨的挑戰是，使用傳統的分析工具來分析密集且不均勻的資料是很麻煩的。資料中可能隱藏著一個結構，但很難找到。面對雜亂的資料，會覺得平均方法很誘人，但這樣做的話會使模式隱藏起來，並使得結果無法代表任何真實的群體。例如，如果你對所有客戶的人口統計資訊進行平均，並且僅僅根據平均值提供一種客戶體驗，那麼你肯定無法聚焦到大多數社群：因為社群傾向於圍繞年齡、職業或婚姻狀況和位置等相關因素聚集。

碁峰 www.gotop.com.tw

此外，對於動態行為，特別是在突發事件和突發事件引發的漣漪，無法透過一時靜態狀態看出來。舉例來說，如果你想像一個社會團體中的人際關係不斷增加，你也會預期溝通的次數更多。這可能導致某次協調破裂和隨後聯盟的組成，或者像在選舉中的子群形成和兩極分化。想預測一個網路隨著時間的發展，所需的方法很複雜，但是如果我們瞭解資料中的結構和相互作用關係，我們可以推斷出動態行為。由於圖形分析著重在關係上，所以能用預測團體韌性。

圖形分析的使用

在最抽象的層次上，圖形分析應用於預測行為和預測變動群組的行動。這樣做需要理解群組中的關係和結構。圖形演算法透過視察網路的連接來實現這一點。透過這種方法，您可以瞭解所連接系統的拓撲結構並對其流程建模。

圖 1-9 中將問題分成三類，用來讓您評估您的工作是否有必要做圖形分析和演算法。

圖 1-9 圖形分析能解答的問題分類。

下面是一些使用了圖形演算法的問題，是否和你面臨的問題相似呢？

- 調查疾病的傳播途徑或是層級傳輸失敗。
- 發現網路攻擊中最易受攻擊或損壞的單位。
- 找出最便宜或最快的資訊或資源路由方法。

碁峰 www.gotop.com.tw

- 預測資料中遺失的連結。
- 確定複雜系統中的直接和間接影響。
- 找出看不見的層次結構和依賴關係。
- 預測團隊會合併還是拆分。
- 找出瓶頸或誰有能力拒絕 / 提供更多資源。
- 根據行為找到群體，以給出適合的個人化建議。
- 減少詐欺和異常檢測中的誤報。
- 為機器學習找到更多預測特徵值。

本章總結

在這一章中，我們看到現今的資料是如何緊密相連的，以及它所隱含的意義。在分析群組動態和關係方面有存在著強大的科學基礎，但這些工具在企業中並不常見。在評估進階分析技術時，我們應該考慮資料的性質，以及是否需要瞭解社群屬性或預測複雜行為。如果我們的資料代表一個網路，我們應該避免將影響因子的數量降低到只剩平均值。相反地，我們應該使用與我們的資料和我們正在尋求的見解相匹配的工具。

在下一章中，我們將介紹圖形概念和術語。

碁峰 www.gotop.com.tw