知識天地

大型網絡之分析

楊振翔助研究員、劉維中助研究員、潘建興助研究員(統計科學研究所)

華嚴經上有一段關於『因陀羅網』的記載：「忉利天王帝釋宮殿， 張網覆上，懸網飾殿。彼網皆以寶珠作之，每目懸珠，光明赫赫，照燭明朗。珠玉無量，出算數表。網珠玲玲，各現珠影。一珠之中，現諸珠影。珠珠皆爾，互相影現。無所隱覆，了了分明。相貌朗然，此是一重。各各影現珠中，所現一切珠影，亦現諸珠影像形體，此是二重。各各影現，二重所現珠影之中，亦現一切。所懸珠影，乃至如是。天帝所感，宮殿網珠，如是交映，重重影現，隱映互彰，重重無盡。」帝釋宮殿中的『天網』上，每一顆寶珠都顯現其它所有寶珠的倒影。而每一個倒影，又顯現其它寶珠倒影的倒影。如此神奇瑰麗的珍寶，是否也曾出現在塵世間呢？

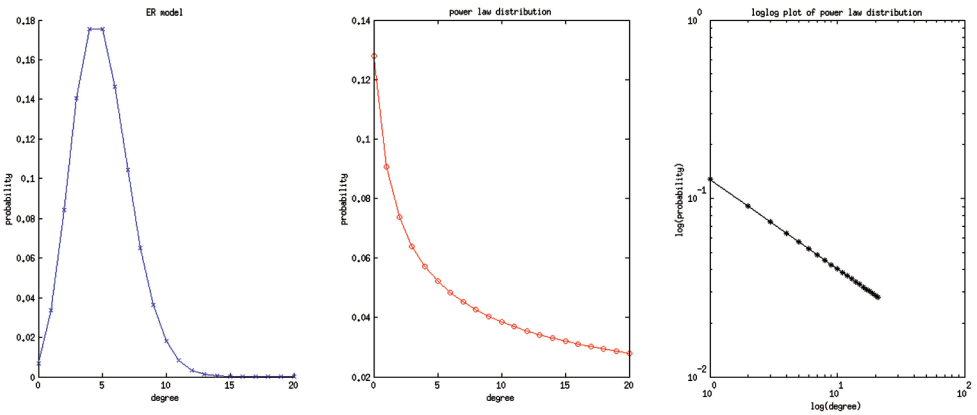
網際網路、臉書、神經系統、基因調控機制、生態系的食物鏈、公司間的生意往來這些分歧而複雜的現象，都是所謂的『大型網絡』。數學上表現網絡關係最簡單的結構是『圖』(graph)。一個graph是由兩種元素所組成：節點(node)和邊(edge)，每一條邊連接兩個節點，此兩節點互為彼此的『鄰居』(neighbor)，而邊可以有方向性或無。例如在社會網絡中，節點代表個人，而邊代表個人間的友誼；在基因調控網絡中，節點代表基因，而邊代表基因間的調控關係。小型的網絡可以經由視覺化的處理，由肉眼來判定其聯結模式。然而大型網絡動輒成千上萬(甚至上億)個節點，這時候肉眼就很難讀取任何有用的資訊了。因此，我們需要更好的量化工具來研究分析這些複雜的大型網絡。

回到前文的例子，在『因陀羅網』上，每兩個節點之間都有一條邊相鄰。這樣的連結方式應該不曾出現在我們所知道的任何大型網絡上。那麼大型網絡的聯結模式為何？它產生的機制，又與其聯結的模式有何關係？

讓我們先從一個簡單的例子看起。假設在一個雞尾酒宴會上，所有的賓客都不認識彼此。在正常情況下，每個賓客花在與一個人交談的時間有一固定範圍，因此在宴會結束時，任意兩個人交談過的機率是一個常數p。如此產生的網絡，稱為『隨機圖』(random graphs)或Erdos-Renyi模型(Erdos-Renyi model, 簡稱ER模型，以最初提出random graphs的兩位匈牙利數學家命名)。嚴格說來，ER模型是網絡的機率分布，而非某一特定的網絡。在雞尾酒宴會的例子中，每一個賓客是一個節點，而兩個交談過的賓客會形成一條邊。

在圖論(graph theory)中一項重要的量是節點的聯結數(connectivity或degree)，也就是和一個節點以邊直接相連的其它節點數目。雞尾酒宴會的例子裡，一位賓客的聯結數是他交談過其他賓客的數目。在ER模型中，這個量是一個隨機變數(random variable)，因此我們可以統計它的機率分布。

圖一左的曲線是一個卜松分布(Poisson distribution)。與常態分布相似，絕大部分的機率都集中在平均值(mean)上，假使聯結數稍微偏移了平均值一些，機率馬上顯著地往下掉。在雞尾酒宴會的例子裡，卜松分布符合我們的直覺猜測：絕大多數賓客交談過的人數是固定的，特別活躍或自閉的情形之機率非常低，以致於出現的數目通常為零。



圖一：左：ER網絡連結數分布，中： 冪次現象網絡連結數分布，右：冪次現象網絡連結數分布的對數圖。橫軸為連結數，縱軸為機率。

然而真實世界大型網絡的聯結數分布並非卜松分布，而是一個遞減函數(圖一中)。如果我們將橫軸(聯結數)與

縱軸(機率)的數值取對數，就會發現兩者呈現線性關係(圖一右)。以數學式表示，令x為聯結數，y為機率，則

$$y = Cx^{-r}, \log y = \log C - r \log x \qquad (1)$$

具有方程式(1)(圖一中) 的聯結數分布之網絡遵循所謂的『 冪次現象 』(power law)，大多數目前我們觀察到的大型網絡，一般而言都遵循乘冪律。在這類網絡中，聯結數越高的節點數目越稀少。大多數的節點都只有一或兩條邊聯結到鄰居，但卻存在少數幾個聯結數非常高的『中心』(hub)，它們可能和成千上萬的節點直接相鄰。在大型網絡中不難找到這些中心節點的身份，例如全球通訊網(World Wide Web)上的www.google.com或www.wikipedia.org，基因調控網絡上的TP53，『推特』(twitter)上的歐巴馬等等。比較ER模型與乘冪律網絡的聯結數分布(圖一左與中)，我們可以很容易看出它們的顯著不同。ER模型的節點非常『平等』，因為所有節點的聯結數幾乎都一樣。相反的，乘冪律網絡是一個很不平等的世界，越重要(聯結數高)的節點數目越少。

冪次現象的機率分布對統計學、自然、社會科學家而言並不陌生。在經濟學上，個人的所得遵循『巴瑞圖分布』(Pareto distribution，一種power law的機率分布)，少數人掌握了社會上大多數的財富。在語言學上，每個字(或詞)在對話或文章中出現的頻率遵循『Zipf's law』(也是一種power law的機率分布)，少數字詞出現的機率非常高(例如中文的『的』或英文的『the』)，而大多數字詞可能只出現一兩次而已。在不同的領域，對乘冪律分布產生的機制有不同的解釋，那麼在大型網絡中，是什麼機制造成聯結數的power law機率分布呢？

匈牙利裔美籍物理學家巴爾巴西(Albert-Laszlo Barabasi)提出了『成長與偏好連結』(growth and preferential attachment)的模型來解釋網絡聯結數的乘冪律。在此一模型中，網絡上節點的數目會隨時間而增加。當新的節點加入網絡時，會偏好與原先聯結數高的節點建立新的邊。這種『富者越富』或『高者越高』的過程，就會讓一開始具有些微優勢的節點最後演變為中心。

並不是所有的社會網絡都具有冪次現象的特性。例如臺灣最大的BBS社群PTT，這個龐大的線上社群提供了一個讓使用者在不同的討論版上討論內容的媒介。PTT可以被想像成一個社會網絡，其節點是版和使用者，而邊則是討論版和使用者的關係。版的大小取決於使用者在板上留言的次數。當我們把所有討論版的大小分布畫成柱狀圖時，令人意外的是此PTT社會網絡並無冪次現象的特性，於是1999年Barabasi and Albert所提出的網絡產生機制：成長與偏好連結或許在這裡是需要修正的。

以一個標準的論壇網頁作例子，當一個使用者到達這個網頁的首頁，正準備要選擇他要去的討論版時，假設他事先並沒有參加特定的討論版，那麼在正常的情況下，他應該會選擇一個比較大的討論版。從網絡產生的機制來看，大的討論版會越來越大，而小的討論版會越來越小。在經過一段時間後，版的大小在柱狀圖上會出現冪次現象。

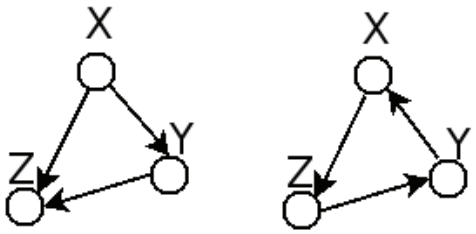
Phoa and Liu (2012)一文提及了自1999年Barabasi and Albert的文章發表後的一些修正機制。其中比較重要的修正是Pennock et al. (2002)的文章。作者把優先連結比喻為勝者完勝(Winner takes all)的現象。他們認為在冪次現象下，在一段很長的時間後，整個論壇網頁只會剩下幾個最大的討論版，而大部分其他的討論板將會無法生存，其大小將會歸零。可是，在現實的網絡世界裡，小的討論版仍然生存，其大小也比優先連結所估計的為大。因此他們在冪次現象之下加上一個均勻分布，用來代表一群支持小的討論版的使用者行為。在這個模型下，版的大小在柱狀圖上會變成一個有長尾現象(long tail)的對數正態分布(lognormal distribution)。

不論是偏好連結，又或是它的修正，這些網絡產生機制都只是根據網絡的結構而產生的。它們都忽略了一些非網絡結構內，但是會對網絡發展有影響的社會和心理因素，Phoa and Liu (2012)稱其為隱藏因素(hidden attributes)。例如，在論壇網頁裡，討論版的大小會被使用者的興趣和投入程度，討論版的內容覆蓋範圍等非網絡結構的因素所影響。Phoa and Liu (2012)提出了兩個方法，其原理大致是先行把相似特性的討論版分門別類，然後每個類別的討論版都以不同的模型去解釋，最後再將所有模型連結在一起。這種分門別類的方法有效減低隱藏因素對網絡產生的影響，而這個模型對討論版大小的估計也比之前的模型更精確。

另外一個可行的網絡模型是完全擺脫偏好連結這個網絡生長的模式，甚至從反方向思考。此模型是建立於一個簡單的機制：相較於一個小群體，一個很直覺的想像是一個大的群體包含了各式各樣的人，因此意見也較多元

而也比較有分裂的張力，來產生較小的討論版群。比如說起初PTT只有一個很大的討論版，於是全部的使用者都連結到這個版，然而隨著時間的增長，不同的使用者們將從大的討論版分裂出來再聚集於較小的討論板，然後以此類推而產生各式各樣不同的PTT討論版群。而此簡單的機制是一種網絡分裂的模型，而其演化所產生的PTT討論版的連結數的分布圖是很接近實際所觀察的PTT社會網絡的。

網絡分析在社會網絡研究裡的應用是很廣泛的。以下是個頗有趣的例子：我們都知道人與人的關係除了朋友關係外也可包含其他的關係類別，比如說，敵意、尊重、倚賴、影響等等。所以不同的人與人之間的連結屬性會產生不同種類的社會網絡。在社會心理學裡，有一個傳統的研究議題是研究三個人i, j和k的人際關係：比方說假如i和j是朋友，那他們是否對k持有相同的看法與態度；相反地，假如i和j是敵人，那他們是否對k持有不同的看法與態度。從另一個角度來說，這個議題就相同於類似我們的人生經驗：朋友的朋友是否是我的朋友，而朋友的敵人是否是我的敵人；或者是往更複雜的角度去想，我的敵人的敵人到底是我的朋友還是敵人，等等的人際關係的問題。一個蠻經典的人際關係的研究是Samuel F. Sampson 的1968年的博士論文。Sampson在美國的新英格蘭的一間修道院紀錄了人與人之間的8個不同的人際關係，其中4個為正向(如誰喜歡誰)，而4個為負向的關係(如誰不尊敬誰)。那我們就以Sampson的資料來回答以上的議題。在此我們用社會網絡分析的『關係代數』(relation algebra)來探討。簡單來說，A、B、C皆是矩陣，分別代表人與人之間不同的關係，那麼A×B=C這個算式就代表A關係經過B關係，變成C關係。假如A、B、C代表友誼網絡，那這個算式的背後意義就代表朋友的朋友會不會是朋友。又或者A是朋友關係，B與C是敵人關係，那A×B=C 就代表朋友的敵人是否是敵人。在Sampson的資料中，總共有8個不同的人際關係，所以在A×B=C的架構下，會有512(=8x8x8)個不同意義的算式。然後我們再用統計的手法來檢定每一個算式是否有統計上的意義，也就是說，A×B所得到的矩陣是否與C矩陣類似，假如類似，則A×B=C這個算式是成立。在512個算式中，若只考慮關係的正負號，那些被驗證而成立的算式，大部分皆屬於以下幾個類別：正正得正，正負得負，負正得負，負負得正。具體來說，當兩個人的關係是正向的話，那他們對第三者的看法也較傾向同時是正的，或同時是負的；相反的，若兩人的關係是負向的話，那他們對第三者的看法，則較傾向為對立的，即一正一負。這個發現也對應了傳統社會學的結構平衡(structural balance)理論，也就是說在一個團體裡人際關係雖有正有負，不過人際關係在社會網絡裡的配置往往會傾向減低衝突發生的可能性。

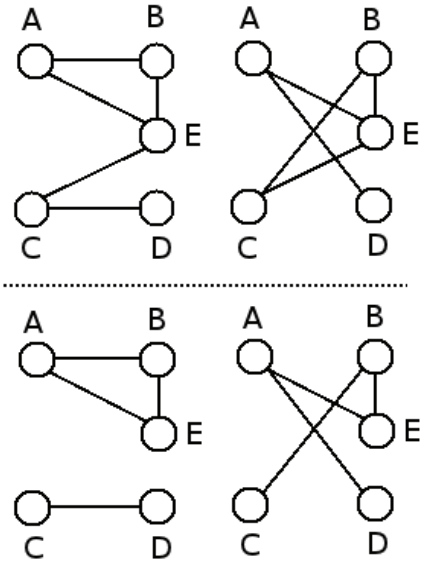


圖二：左：前饋迴路，右：反饋迴路。

冪次現象屬於網絡的『巨觀』(macroscopic)性質。除此之外，大型網絡也具有獨特的『微觀』(microscopic)性質，例如『網絡母題』(network motif)。

在文學或藝術上，『母題』(motif)代表反覆出現的『模式』(pattern)，例如貝多芬命運交響曲的前四個音。此一概念首先被以色列生物學家阿隆(Uri Alon)應用於計算生物學或系統生物學上。 在網絡上， 模式代表一小組節點聯結的拓樸結構。例如，考慮所有三個節點連在一起(connected)的無方向性網絡，總共只有兩種情形：一個串一個的『鏈』(chain)或兩兩互鄰的『小集團』(clique)。拓樸模式的數目隨節點數目的增加而呈指數性的增長。

Alon發現在生物網絡中，有些拓樸模式出現的頻率要比其它拓樸模式高得多。他比較兩種三個節點的拓樸模式：『前饋迴路』(feed forward loop)與『反饋迴路』(feed back loop)(圖二左與右)，在大腸桿菌(Escherichia coli)基因調控網絡中出現的頻率，發現前者比後者的頻率高得多。



圖三：上：經過邊交換後，三個節點母題數目不變，下：經過邊交換後，三個節點母題數目改變。

何以某些拓樸模式會成為生物網絡的母題，而其它則否？他認為這些拓樸模式在基因調控中有某些演化上的優點。例如前饋迴路可以讓基因對環境的反應更穩定，不受高頻雜訊(shot noise)的影響。然而Alon對網絡母題的探討，大多限於生物系統上。同樣的概念是否能應用在其它大型網絡上？

這是我們對於網絡母題的研究動機，具體而言有三個目標：

1. 建立量化網絡母題的統計方法。

2. 找尋不同種類網絡中的母題。

3. 詮釋這些母題在不同網絡中的意義。

如前所述，網絡母題是在網絡上反覆出現的拓樸結構。如何讓這個模糊的描述變得精確？在統計上，當我們宣稱某事件出現次數是有意義的，通常是根據觀察的結果否證一個『虛無假設』(null hypothesis)。例如要證明某藥物對某疾病是否有療效，就必須比對服藥病人與服用安慰劑病人的反應。此處的虛無假設是『服用該藥與服用安慰劑的效果並無差異』。那麼什麼是判定網絡母題的虛無假設呢？

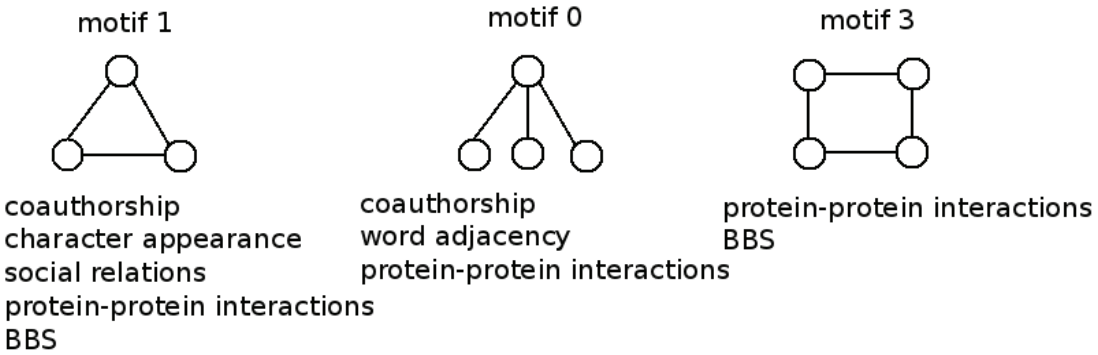
這個虛無假設應是隨機抽樣產生網絡的機制，而隨機抽樣出來的網絡，應該要和實際觀察到的網絡越接近越好。因此必須符合下列條件：

1. 抽樣網絡的節點數與邊數與實際觀察到的網絡相同。

2. 抽樣網絡的連結數分布與實際觀察到的網絡相同。

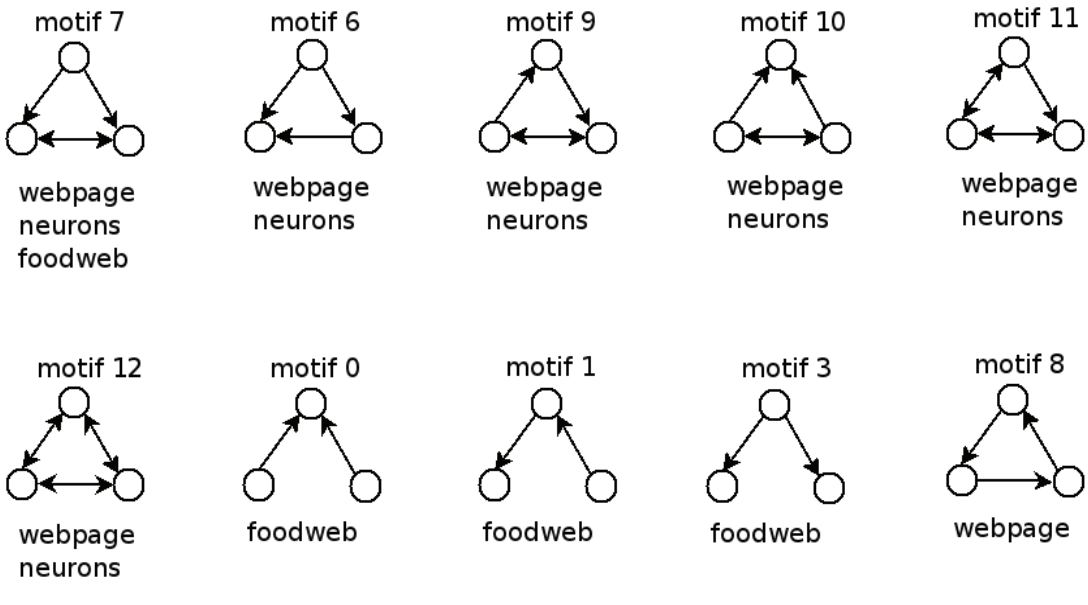
3. 抽樣網絡不改變『低階』(lower order)母題的數目。

條件1與2可以經由交換兩條邊上的節點(edge swap)達到：把兩條邊A-B與C-D改成A-D與B-C。這樣的交換不會改變任何節點的連結數(只是連結節點不一樣了)，當然更不會改變網絡的大小。我們可以隨機進行邊交換，直到所有的節點至少被交換了一次為止。



圖四：無方向性網絡上出現的三節點與四節點的母題。

然而邊交換還不是建構網絡母題虛無假設的充分條件。當我們宣稱某一網絡母題出現頻率在統計上是有意義時，我們必須把其它相關低階網絡母題的效應考慮進去。因此，在考慮四個節點的母題時，我們只容許不改變三



圖五：有方向性網絡上出現的三節點的母題。

個節點母題數目的邊交換。例如圖三上邊的網絡經過邊交換後，三個節點母題數目不變，然而下邊的網絡經過邊交換後，會減少一個三角形並增加兩個鏈。所以只有左邊的網絡邊交換是被允許的。

我們利用此種方法尋找十多個網絡上的母題(Yeang, Huang and Liu 2012)。這些網絡包括了(1)網站的連結(2)神經系統、蛋白質交互作用等生物網絡(3)生態系的食物網(4)科學家是否有共同著作的關係(5)社交網絡(6)英文單字在句子中是否相鄰(7)小說人物是否出現在同一章節(8)BBS用戶是否回應其他用戶的文章。

圖四顯示在無方向性網絡上出現的三節點與四節點的母題。三角形(母題1)出現在大多數的網絡上，這表示大多數網絡關係都有遞移性(transitivity)，例如在社交網絡中，朋友的朋友通常也是朋友。少數例外之一是英文單字的相鄰，因為語言是線性的，所以AB相鄰與BC相鄰時，通常AC不會相鄰。

星形結構(母題0)也是常出現的母題。這類網絡存在一些中心(hub)，像是英文單字中的the，或小說的主角人物，其它節點與中心相鄰但彼此不相鄰。此外，鑽石結構(母題3)出現在蛋白質交互作用與BBS網絡上，這類母題出現的機制並不清楚。

圖五顯示在有方向性網絡上出現的三節點的母題。有趣的是，在食物網出現的所有母題，都據有清楚單一的方向性，例如母題0，1或3。這是因為大多數物種的食性都有單一方向性，例如草食動物不會吃肉食動物。反之，網站間的連結會有更複雜的模式，例如 反饋迴路(母題8)。

大型網絡分析已成為目前的顯學，不論在理論上或實務上都有許多重要的問題尚待解決。這些問題具有強烈的跨領域性，因此吸引了各學門的人來鑽研。在受不同訓練的心智激盪下，讓這些問題變得更有趣了。

參考文獻

1. Barabasi A.L., Albert R. (1999). Emergence of Scaling in Random Network. Science 286: 509-512.
2. Pennock D.M., Flake G.W., Lawrence S., Glover E.J., Lee G. C. (2002). Winners Don't Take All: Characterizing the Competition for Links on the Web. Proceedings of the National Academy of Sciences 99: 5207-5211.
3. Phoa F.K.H. and Liu W.C. (2012). High-Quality Winners Take More: Modeling Non-Scale-Free Bulletin Forums with Content Variations. (submitted).
4. Sampson S (1969). Crisis in a cloister. Unpublished doctoral dissertation, Cornell University.
5. Wasserman S., Faust K. (1994). Social network analysis: methods and applications. Cambridge University Press.
6. Milo R., Shen-Orr S., Itzkovitz S., Kashtan N., Chklovskii D., Alon U. (2002). Network motifs: simple building blocks of complex networks. Science 298:824-827.
7. Yeang C.H., Huang L.C., Liu W.C. (2012). Recurrent structural motifs reflect characteristics of distinct networks. Proceedings, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).