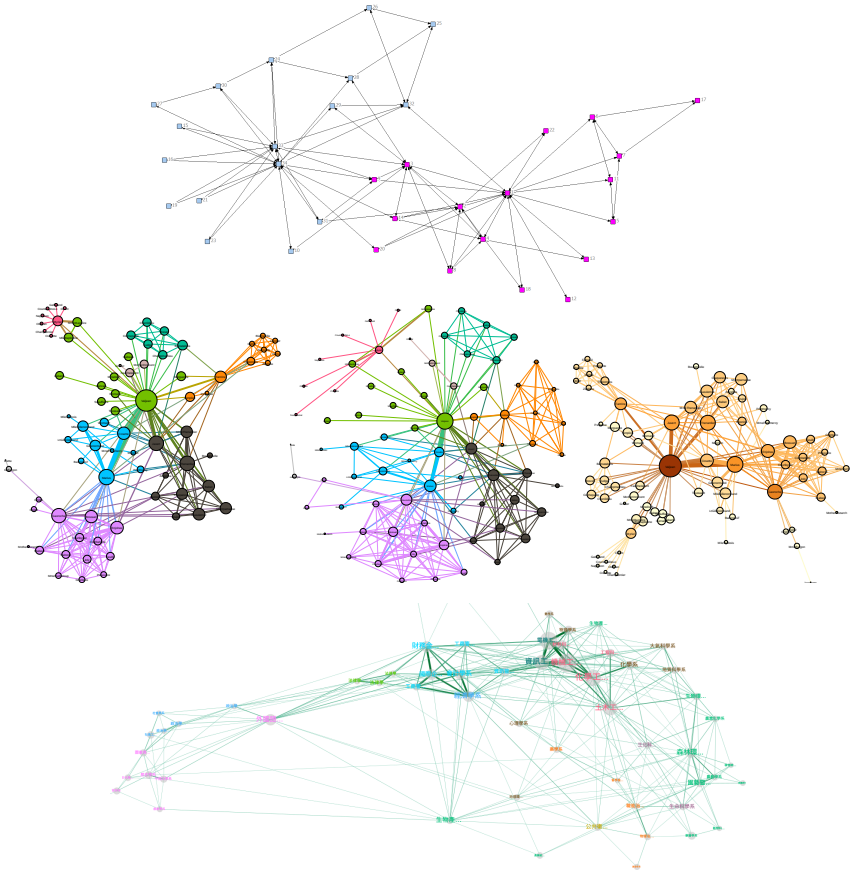
# 社會網絡分析專題：作業一



# 目錄

# 1

# 第一題

## 1.1 Support Attribute

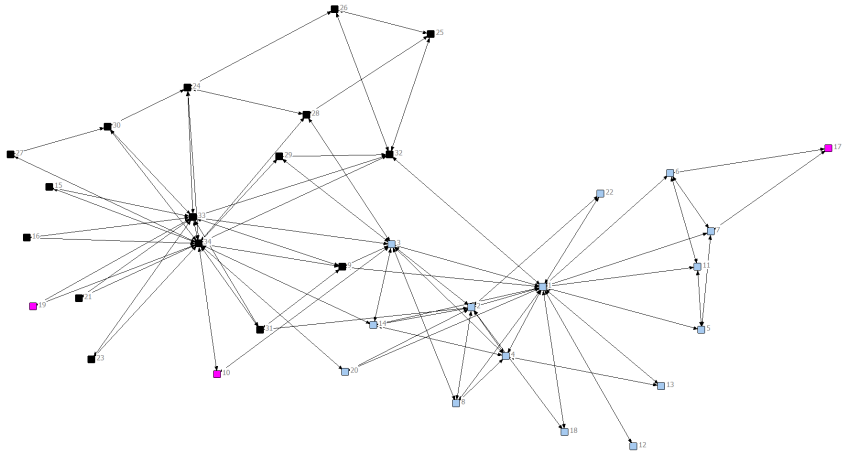
### 1.1.1 network-level

| 項目 | 數值 | 說明 |
|---|---|---|
| E-I Index | -0.615 | 內向性 > 外向性 |
| Expected value | 0.187 | |
| Permutation Test p ⩽ Ob | 0.000 | p < 0.025時，E-I index值顯著 |

在計算網絡整體的凝聚力(cohesion)時，因為E-I index為負數，因此得知此網絡較偏內向性。經由排列檢定(Permutation test)檢驗E-I index顯著與否時，由於在信心水準95%的雙尾檢定下，若p值小於0.025則表示E-I index值顯著，本題的E-I index(-0.615)小於期望值(0.187)，需檢視 p ⩽ Ob(0.000)之欄位，故可判斷此網絡整體的內向性是顯著的。

### 1.1.2 group-level

| | E-I Index | 說明 |
|---|---|---|
| **group 0** | 1.000 | 完全外向 |
| **group 1** | -0.676 | 內向性 > 外向性 |
| **group 34** | -0.684 | 內向性 > 外向性 |



如上圖所示，該網絡在此屬性下可分為三種小團體：

group 0所得E-I index(1) 為完全外向，表示組內的成員彼此完全不會聯繫，皆與另外兩組成員交流。group 1之E-I index(-0.676) 與group 34之E-I index(-0.684) 內向性高，表示組內成員與組內成員之連結大於與組外成員的連結。從視覺化的圖可看出group 1與group 34分別是支持網絡內成員編號1與34的小團體，goup 0之成員則可視為未表態、零散的網絡內個體。

### 1.1.3 individual-level

### 1.1.3 individual-level

| | | Intern | Extern | Total | E-I |
|---|---|---|---|---|---|
| group 0 | | | | | |
| | 10 | 0.000 | 2.000 | 2.000 | 1.000 |
| | 17 | 0.000 | 2.000 | 2.000 | 1.000 |
| | 19 | 0.000 | 2.000 | 2.000 | 1.000 |
| group 1 | | | | | |
| | 3 | 5.000 | 5.000 | 10.000 | 0.000 |
| | 20 | 2.000 | 1.000 | 3.000 | -0.333 |
| | 6 | 3.000 | 1.000 | 4.000 | -0.500 |
| | 7 | 3.000 | 1.000 | 4.000 | -0.500 |
| | 14 | 4.000 | 1.000 | 5.000 | -0.600 |
| | 2 | 8.000 | 1.000 | 9.000 | -0.778 |
| | 1 | 14.000 | 2.000 | 16.000 | -0.750 |
| | 4 | 6.000 | 0.000 | 6.000 | -1.000 |
| | 5 | 3.000 | 0.000 | 3.000 | -1.000 |
| | 8 | 4.000 | 0.000 | 4.000 | -1.000 |
| | 11 | 3.000 | 0.000 | 3.000 | -1.000 |
| | 12 | 1.000 | 0.000 | 1.000 | -1.000 |
| | 13 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 18 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 22 | 2.000 | 0.000 | 2.000 | -1.000 |
| group 34 | | | | | |
| | 9 | 3.000 | 2.000 | 5.000 | -0.200 |
| | 29 | 2.000 | 1.000 | 3.000 | -0.333 |
| | 28 | 3.000 | 1.000 | 4.000 | -0.500 |
| | 31 | 3.000 | 1.000 | 4.000 | -0.500 |
| | 34 | 13.000 | 4.000 | 17.000 | -0.529 |
| | 32 | 5.000 | 1.000 | 6.000 | -0.667 |
| | 33 | 10.000 | 2.000 | 12.000 | -0.667 |
| | 15 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 16 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 21 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 23 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 24 | 5.000 | 0.000 | 5.000 | -1.000 |
| | 25 | 3.000 | 0.000 | 3.000 | -1.000 |
| | 26 | 3.000 | 0.000 | 3.000 | -1.000 |
| | 27 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 30 | 4.000 | 0.000 | 4.000 | -1.000 |

**group 0**

five most out-ward: 10, 17, 19。僅有的三位成員皆為外向，完全不會與組內另外二人聯繫。

**group 1**

five most out-ward: 3, 20, 6, 7, 14。除了3 (0)與組內外聯繫平均之外，其他四者皆為負值。

five most in-ward: 4, 5, 8, 11, 12, 13, 18, 22。這些人之E-I index值皆為-1。

**group 34**

five most out-ward: 9, 29, 28, 31, 34。

five most in-ward: 15, 16, 21, 23, 24, 25, 26, 27, 30。這些人之E-I index值皆為-1。

## 1.2 Club Attribute

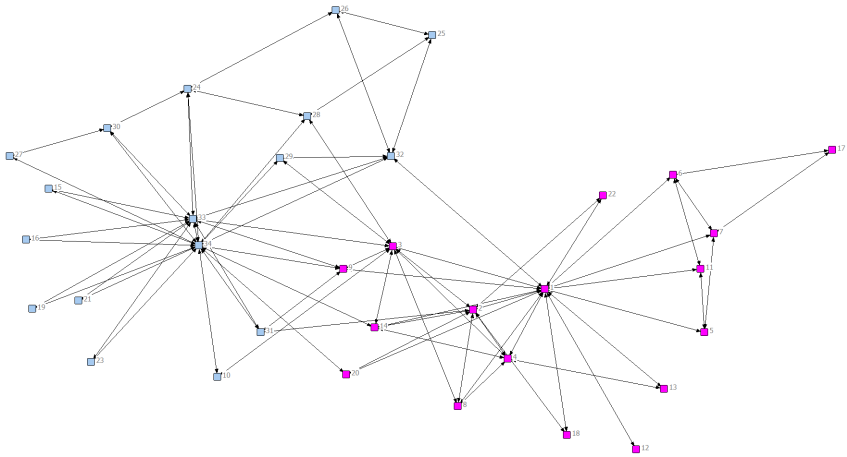
### 1.2.1 network-level

| 項目 | 數值 | 說明 |
|---|---|---|
| E-I Index | -0.718 | 內向性 > 外向性 |
| Expected value | 0.030 | |
| Permutation Test p ⩽ Ob | 0.000 | p < 0.025時，E-I index值顯著 |

使用Club的屬性檢驗時，E-I index為負數（偏內向性），並且 p ⩽ Ob 欄位之值(0.000)小於0.025，故網絡整體的內向性同樣是顯著的。

### 1.2.2 group-level

| | E-I Index | 說明 |
|---|---|---|
| **group 1** | -0.728 | 內向性 > 外向性 |
| **group 34** | -0.707 | 內向性 > 外向性 |



如上圖所示，該網絡在此屬性下可分為兩種小團體，且兩組內部聯繫皆遠熱絡於外部聯繫，故E-I index皆為負數。

### 1.2.3 individual-level

### 1.2.3 individual-level

| | | Intern | Extern | Total | E-I |
|---|---|---|---|---|---|
| group1 | | | | | |
| | 9 | 2.000 | 3.000 | 5.000 | 0.200 |
| | 3 | 6.000 | 4.000 | 10.000 | -0.200 |
| | 20 | 2.000 | 1.000 | 3.000 | -0.333 |
| | 14 | 4.000 | 1.000 | 5.000 | -0.600 |
| | 2 | 8.000 | 1.000 | 9.000 | -0.778 |
| | 1 | 15.000 | 1.000 | 16.000 | -0.875 |
| | 4 | 6.000 | 0.000 | 6.000 | -1.000 |
| | 5 | 3.000 | 0.000 | 3.000 | -1.000 |
| | 6 | 4.000 | 0.000 | 4.000 | -1.000 |
| | 7 | 4.000 | 0.000 | 4.000 | -1.000 |
| | 8 | 4.000 | 0.000 | 4.000 | -1.000 |
| | 11 | 3.000 | 0.000 | 3.000 | -1.000 |
| | 12 | 1.000 | 0.000 | 1.000 | -1.000 |
| | 13 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 17 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 18 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 22 | 2.000 | 0.000 | 2.000 | -1.000 |
| group34 | | | | | |
| | 10 | 1.000 | 1.000 | 2.000 | 0.000 |
| | 31 | 2.000 | 2.000 | 4.000 | 0.000 |
| | 29 | 2.000 | 1.000 | 3.000 | -0.333 |
| | 28 | 3.000 | 1.000 | 4.000 | -0.500 |
| | 34 | 14.000 | 3.000 | 17.000 | -0.647 |
| | 32 | 5.000 | 1.000 | 6.000 | -0.667 |
| | 33 | 10.000 | 2.000 | 12.000 | -0.667 |
| | 15 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 16 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 19 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 21 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 23 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 24 | 5.000 | 0.000 | 5.000 | -1.000 |
| | 25 | 3.000 | 0.000 | 3.000 | -1.000 |
| | 26 | 3.000 | 0.000 | 3.000 | -1.000 |
| | 27 | 2.000 | 0.000 | 2.000 | -1.000 |
| | 30 | 4.000 | 0.000 | 4.000 | -1.000 |

**group 1**

five most out-ward: 9, 3, 20, 14, 2。

five most in-ward: 4, 5, 6, 7, 8, 11, 12, 13, 17, 18, 22。這些人之E-I index值皆為-1。

**group 34**

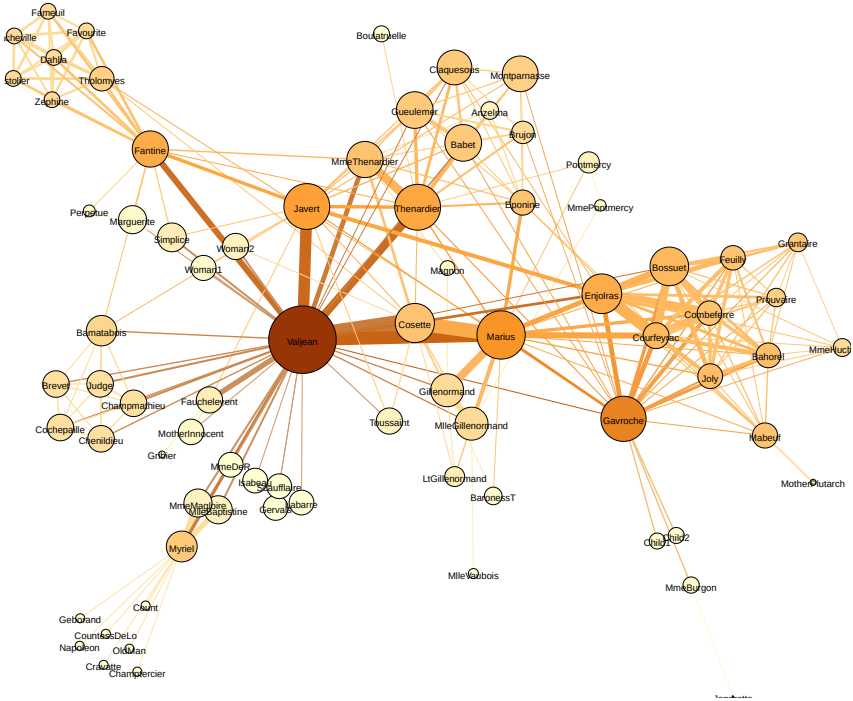
five most out-ward: 10, 31, 29, 28, 34。

# 2

# *第二題*

### 2.1 2.a

- 透過視覺化的方式來呈現 Les Miserables 來呈現角色的網絡關係圖，每一個節點的顏色代表該節點 degree 值的大小，也就是該節點和多少個其他節點連接。每一個節點的大小代表該節點 closeness centrality 值的大小，也就是該節點到其他節點所需要的最短距離總和的倒數，closeness 值越大代表該節點和其他節點越靠近，處於網路中越中心的地方。
- 從圖一中可以發現 Valjean 是最重要的節點，根據節點的大小和深淺也可以判斷 Javert, Thenardier, Marius, Gavroche 也是重要的幾個節點。
- 從圖一中也可以發現由 Fantine 和 Myriel 所延伸出去的節點最不重要，這些節點不僅較小，顏色也較淺，代表和較少的節點連接，並且處於網路的邊陲。



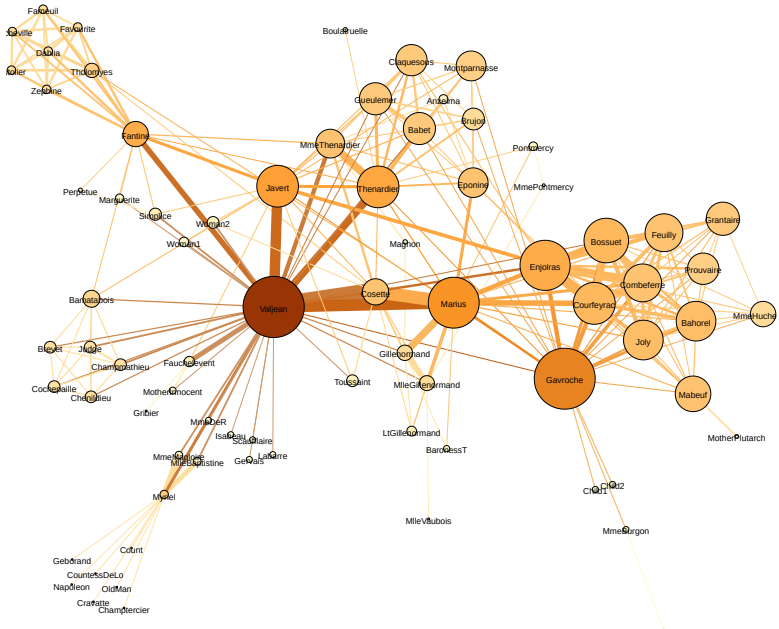

圖一

---

## 2.2 2.b

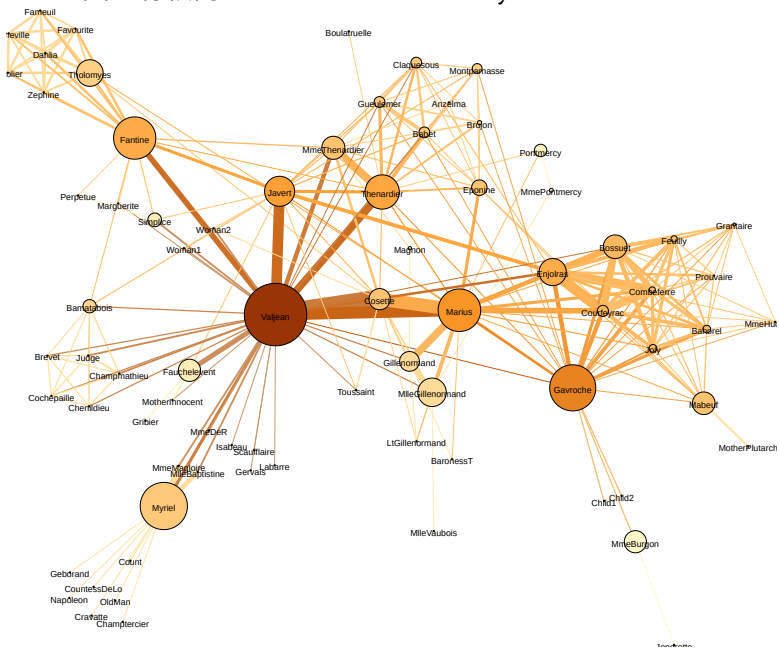
- 瞭解節點的重要性或者性質，除了在 2.a 所使用的 degree 之外，也有其他的指標可以參考。
    - eigenvector centrality：這個指標用以測量一個節點的影響力，若該節點所連接到的其他節點的 eigenvector centrality 高，則該節點會有相對較高的 eigenvector centrality 值。這個指標不僅考慮到節點的 degree 數量，亦考慮了和哪

從圖二來看，顏色依然代表該節點 degree 值的大小，而節點大小則代表該節點 eigenvector centrality 的大小，和圖一相比，圖二由 Marius 和 Gavroche 延伸出去的 cluster 的節點大小變得稍大，而其他肉眼可辨的 clusters 的節點大小則變得較小，尤其是在網路左側的 clusters 節點縮小幅度更大。改為使用 eigenvector centrality 後產生的變化凸顯出那些變大或是沒有什麼變化的節點所連接到的其他節點同樣都是中心性較強的節點（剛好是 2.a 所提到的 Valjean, Javert, Thenardier, Marius, Gavroche 五個重要節點），而其他明顯變小的節點則只有一個或是沒有中心性較強的節點和它們連接。




- betweenness centrality：這個指標用於衡量一個節點是否位於任兩個其他節點之間的最短距離之上，若該節點的 betweenness centrality 高，代表該節點屬於一種橋樑性的

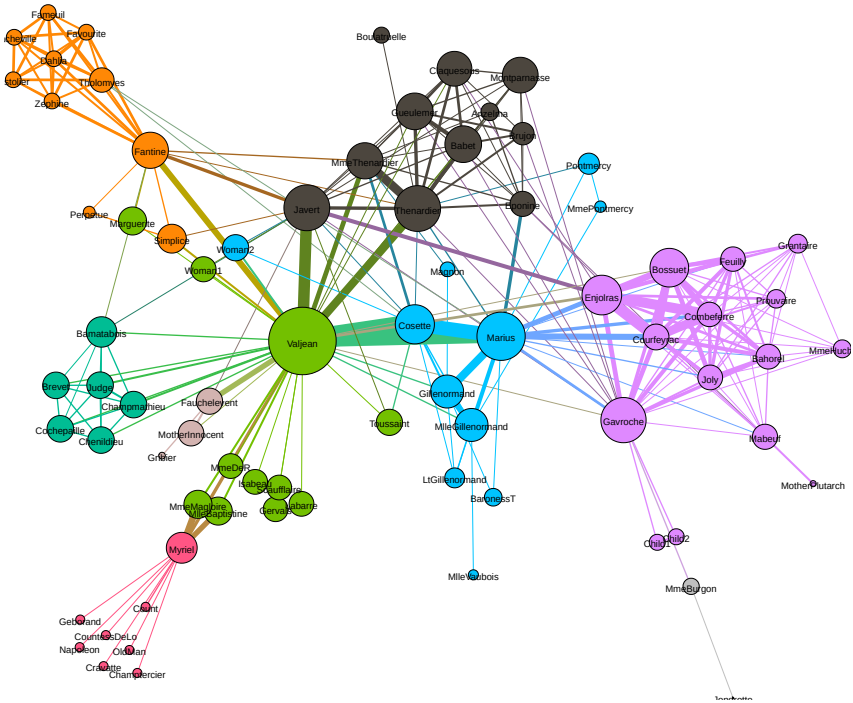
節點大小則代表該節點 betweenness centrality 的大小，和圖一圖二相比，最明顯的差異在於網路左側的 Fantine 和 Myriel 的節點大小被放大很多，在前兩張圖當中，因為這兩個點並不處於中心性的位置（closeness centrality 低），其向外所連接的 cluster 的重要性也不強（eigenvector centrality 低），導致這兩個節點的重要性無法被凸顯出來。實際在網絡中，這兩個節點獨自連接出了數個較為邊陲的節點，使得這些邊陲的節點們必須高度仰賴於這兩個節點才能夠散播或接收到資訊，在這個角度之下，由於 Fantine 和 Myriel 掌控了邊陲節點向外連接的最短距離，因此有較高的 betweenness centrality。




圖三

- 除了節點的特徵之外，網路之中不同 clusters 的分群也是我們關心的特徵，透過分群可以很輕易地判斷出哪些節點之間的同質

推論。例如，若一個節點獨立連接到一個邊陲的群，則可以推論該節點可能有較高的 betweenness centrality，或是若一個節點所連接到的群有較靠近中心的位置且該群之間的 density 很高，則可以推論該節點可能有較高的 eigenvector centrality。分群的結果如圖四所示，顏色代表不同的群體，而節點大小代表 betweenness centrality 的大小。

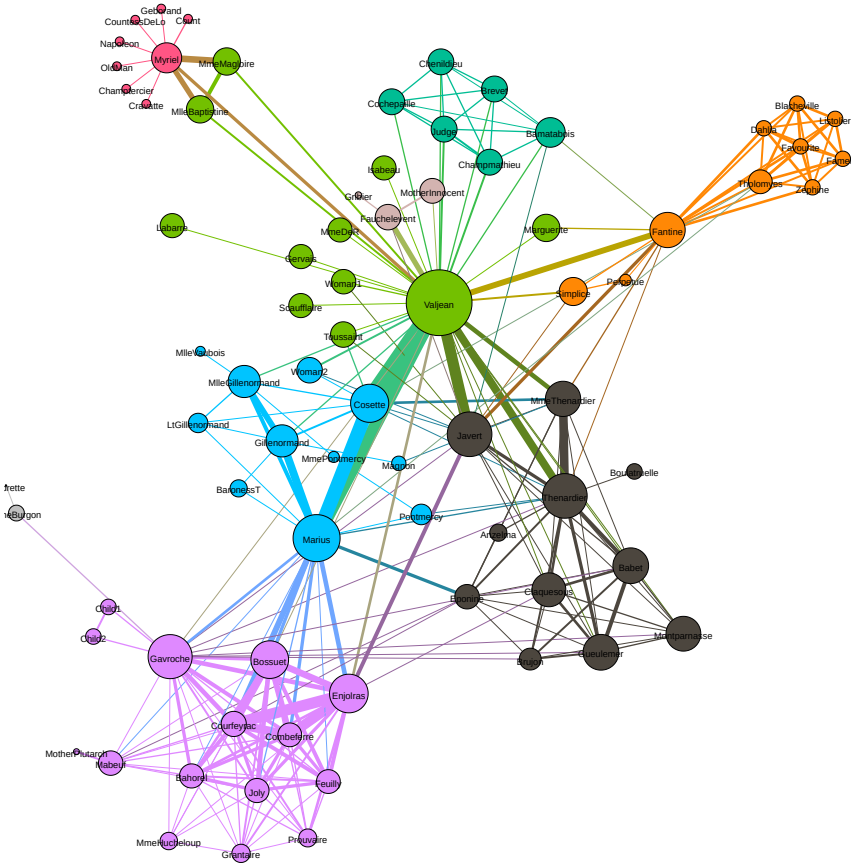



圖四

## 2.3 2.c

本題嘗試利用不同的 layouts 來凸顯網路的特徵，欲呈現的特徵包含節點的 betweenness centrality，透過節點的大小來呈現，節點的分群，透過節點的顏色來呈現，以及節點之間連結的程度，透過連結的粗細來

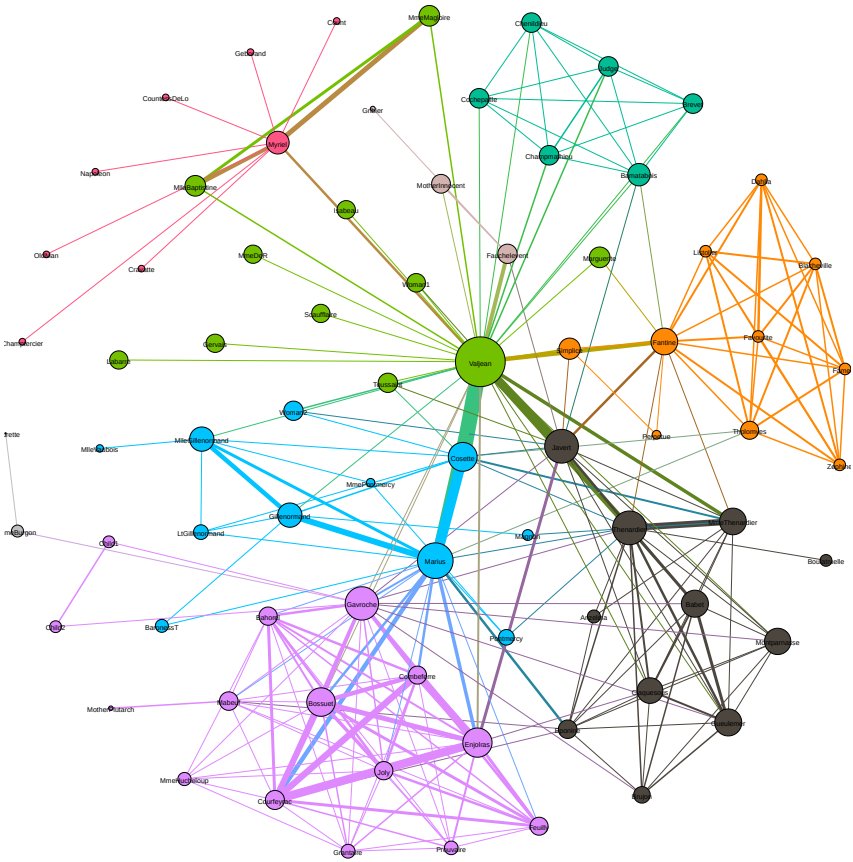
呈現。
- Force Atlas：呈現如圖五所示，不同的 clusters 在位置上被區隔出來，節點的大小也便於區辨其重要性，不同 clusters 的稠密程度也可以從不同顏色的連結的粗細觀察。




圖五

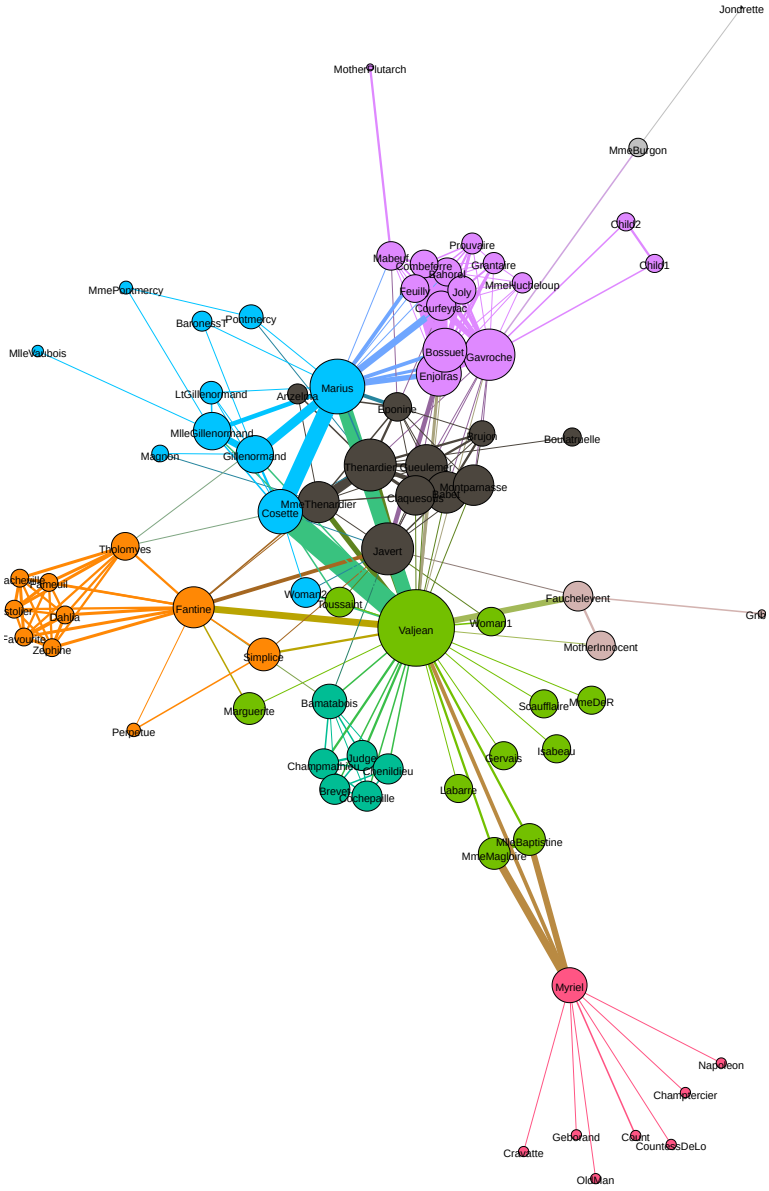
- Fruchterman Reingold：呈現如圖六所示，網路被編排為圓形，不同 clusters 的區別和其各自的稠密程度依然相當明顯，但是不同的節點的重要性變得不明顯。例如次重要的節點像是 Thenardier, Marius, Gavroche 都不在最重要的的節點 Valjean 旁

邊，即使他們之間有連結，在視覺上看起來卻有兩步三步之遙。




圖六

- YiFan Hu：呈現如圖七所示，該圖和以 Force Atlas 的 layout 製成的圖五很類似，只是把靠近中心的節點排列得更加緊密，把處於邊陲的節點拉得更加遙遠。這種呈現方式可以幫助更明確地找出處於邊陲的節點，但對於處於中心的那群節點來說，由於排列過於緊密，在 clusters 的區別上就沒有前兩張圖來得明顯。


Jondrette
MotherPlutarch
MmeBurgon
Child2
Child1
Prouvaire
Mabeuf
Combeferre
Gueulemer
Grantaire
Feuilly
Joly
MmeHucheloup
Courfeyrac
MmePontmercy
Baroness
Pontmercy
MlleVaubois
Bossuet
Gavroche
Enjolras
Marius
LtGillenormand
Anzelma
Boonmie
MlleGillenormand
Brujon
Babet
Magnon
Gillenormand
Thenardier
Claquesous
Montparnasse
MmeThenardier
Cosette
Javert
Tholomyes
Fantine
Woman2
Toussaint
Valjean
Woman1
Fauchelevent
Gnr
MotherInnocent
Simplice
Perpetue
Marguerite
Babatabbis
Scaufflaire
MmeDeR
Isabeau
Gervais
Judge
Champmathieu
Chenildieu
Brevet
Cochepaille
Labarre
TrialBaptistine
MmeMagloire
Myriel
Napoleon
Champtercier
Gervand
Count
CountessDeLo
Crochette
Otharian

圖七

以上三種 layouts 在視覺呈現上都有各自的優劣之處，但整體來說都適合在本資料中想要凸顯的數個特徵。

# 3

# 第三題：Real World Data

## 3.1 構想

大學的科系百百種，在這麼多科系之中，我們能如何了解這些科系的特質？例如，哪些科系是「相似的」？熱門科系有哪些？冷門科系又有哪些？此外，有所謂「特殊的」、難以用前面幾種用語簡單描述的科系嗎？

大學學測申請入學或許提供一種方式，讓我們可以了解大學各科系的特質。這些科系能夠透過申請者形成一個網絡 – 若同一個申請者同時申請了兩個科系，那這兩個科系就形成連結。透過許多科系形成的網絡，我們或許能夠嘗試回答上述的問題。

### 3.1.1 目標資料

新鮮人查榜的學測交叉查榜[1]提供了每年大學申請入學的資料。在每個科系的頁面右欄有現成的網絡資料 – 例如，在臺大法律系的 交叉科系分析[2]，可以直接看到同時申請臺大法律系和其它科系的人數。

### 3.1.2 定義網絡

考慮資料取得的容易程度以及伺服器的負擔，我們決定使用** 104 年學測申請入學臺灣大學各系的資料**作為網絡的範圍定義。在此範圍內，網絡的定義如下：

- Nodes

[1] https://freshman.tw/cross

[2] https://freshman.tw/cross/104/006342

- 定義：臺大某科系。

  例如，臺大圖資為一個 node，臺大森林系為另一個 node。

- Edges
  - 定義：若學生 A 同時申請「科系甲」和「科系乙」，則臺大的甲乙系之間形成連結
  - 特性：
    - Weighted：同時申請「科系甲」和「科系乙」的學生**人數**
    - Undirected (無向連結)

---

## 3.2 資料收集

### 3.2.1 網絡資料爬取

我們使用 Scrapy[3] 將臺灣大學各個科系 (共 60 個) 的資料 從新鮮人查榜[4] 爬取下來。爬取的資料包含：

1. 科系名稱
2. 系所資訊
3. 交叉科系分析
4. 區域分析 (依准考證號)

[3] https://scrapy.org

[4] https://freshman.tw/cross/104/001

此外，臺大各學系所屬學院則由臺大課程網[5]以及臺大課號編碼作業說明[6]所提供之資訊取得。

Scrapy 爬蟲的原始碼託管於 GitHub[7]；詳細的變項描述以及資料清理過程記錄於 Jupyter Notebook[8]。

### 3.2.2 網絡資料描述

清理過後的資料包含 60 個學系 (node) 以及 634 條連結 (edge)。由於「護理學系(公費生)」並未與網絡中的其它科系形成連結，在視覺化時將其剔除。因此，下文視覺化的網絡由 59 個 node 和 634 條 edge 組成。

整理成網絡資料格式的資料存放於 `ntuNetwork_edges.csv`[9] 與 `ntuNetwork_attr2.csv`[10]。

## 3.3 預期與假設

我們希望透過這筆資料能得出兩種資訊：(1) 哪些科系是相似的；以及 (2) 在考慮招生人數以及 degree 這兩種資訊下，哪些科系看起來比較特殊。

1. 的基本假設是「人們在選擇要申請哪些科系時，應該會有某種『偏好[11]』使其選擇『相似』的科系」。因此，透過各科系之間的連結數量以及強度，我們或許可自然地將科系進行分群。

[5] https://nol2.aca.ntu.edu.tw/nol/guest/index.php
[6] https://nol2.aca.ntu.edu.tw/nol/guest/課程編碼說明.pdf
[7] https://github.com/liao961120/collegeSNA
[8] https://liao961120.github.io/collegeSNA/ntuNetwork
[9] https://bit.ly/2OygdeK
[10] https://bit.ly/2U236by
[11] 「偏好」可能是個人興趣，也有可能是家人或社會期待，造成人們在選系時，會選擇一群「相似的」科系。

2. 則假設在不考慮各科系的特性下，**若一科系的招生名額越多，其 degree 也應該越大**，因為一科系招生名額越多，通常也有越多人申請，而這些申請者越有機會同時申請其它的科系。我們可以去觀查**哪些科系違反這項直覺**，再去猜想原因為何。

### 3.4 視覺化網絡

依據上述的假設，我們將以下 node 與 edge 的屬性資料對映 (map) 到網絡圖 (圖 3.1) 中的視覺元素：

- Edge Attributes
    - Thickness & Color gradient: Edge weight (共同申請兩系人數)
        - Minimum: 共同申請人數需大於或等於 3 才形成連結 (資料來源限制)
- Node Attributes
    - Node Size: 招生名額
    - Text size: Degree
    - Text Color: 學院

所有的屬性 (除了 Text Color 之外) 皆有數值梯度，並線性對應到視覺屬性的強度上。

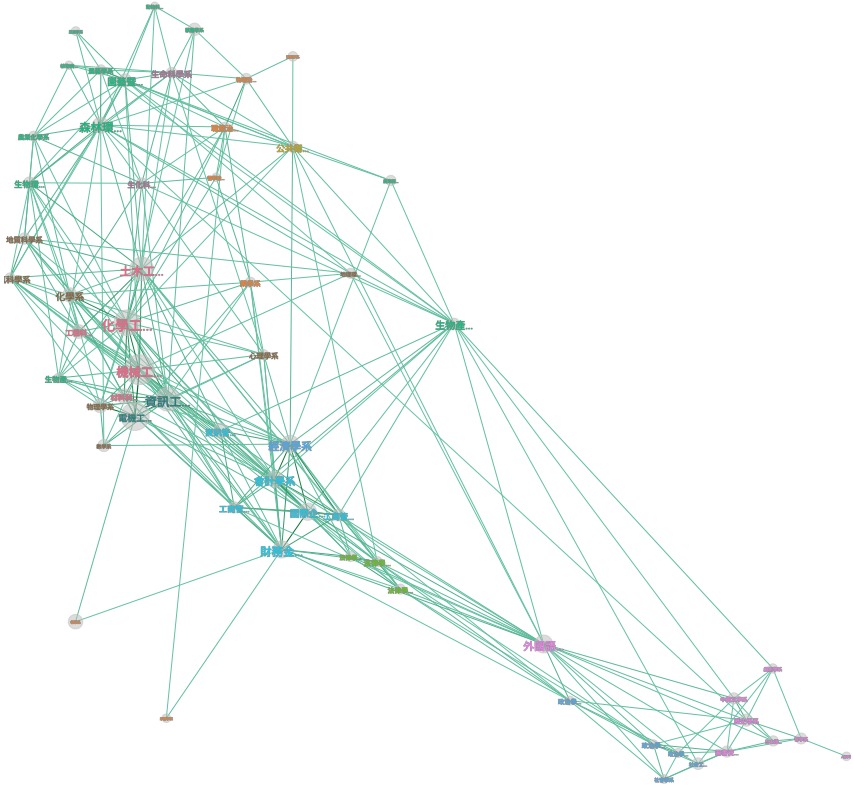



Figure 3.1: **104 年學測申請入學 臺大各系申請網絡連結**。svg 原圖：http://bit.ly/104ntuNetwork

### 3.4.1 Node Size vs. Text Size

根據上述預期與假設的 (2)，我們找出了幾個違反直覺的科系：

- Node 相對 Text 大上許多
    - 意義：申請該系的人不太會申請臺大其它科系
    - 例子：醫學系、人類系、牙醫系 (依偏離程度由大至小排序[12])
- Node 相對

[12]這邊是以視覺直接比較，可能會有誤差。

Text 小許多：申請該系的人也喜歡申請其它科系

- 意義：申請該系的人也會申請臺大其它科系
- 例子：生傳系

由 Node 相對 Text 大上許多的科系，我們或許能做出一些猜想。由於醫學系和牙醫系的申請分數通常很高，其相對較低的 degree 可以反映是**申請者的選擇而非其它因素 (如分數不足) 使其不申請其它科系**。至於人類系的情況則較難詮釋。

至於第二種情況，Node 相對 Text 小許多的情形則比較少見。此情況比較明顯的科系是生傳系。值得注意的是，生傳系在網絡中似乎自成一個 cluster，原因是因為它分別與文學院、管理學院和農學院都有相當數量的連結。

### 3.4.2 分群

這裡視覺化使用 Gephi 的 **MultiGravity ForceAtlas 2** layout algorithm。**MultiGravity ForceAtlas 2** 是與 **ForceAtlas** 效果類似的演算法。關於 **ForceAtlas** 的直覺詮釋如下：

> [I]t simulates a physical system in order to spatialize a network. Nodes repulse each other like charged particles, while edges attract their nodes, like springs. These forces create a movement that converges to a balanced state. (Jacomy, Mathieu AND Venturini, Tommaso AND Heymann, Sebastien AND Bastian, Mathieu 2014)

因此，這個演算法會將彼此**連結較多或較強** 的 Node 放在附近，相當適合用來視覺化分群的結果。

圖 3.1 中的文字顏色代表該系實際上所屬的學院，而圖中空間位置相近的 node 則代表 (由連結的資料上來看) 彼此相似的科系。因此，由圖 3.1 我們可以很快地發現**醫學院是最「鬆散」的學院、理學院其次；工學院則相當緊密，且電資學院與工學院非常相近**。另外一個有趣的現象是，經濟學系與所屬的社科學院脫離，加入以管理學院為主所形成的 cluster。

# 4

# 參考資料

Jacomy, Mathieu AND Venturini, Tommaso AND Heymann, Sebastien AND Bastian, Mathieu. 2014. “ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software.” *PLOS ONE* 9 (6). Public Library of Science: 1–12. https://doi.org/10.1371/journal.pone.0098679.