社會模擬工作坊

Team 5 - 擬會作社模 (負責論文9, 16)

9. How Network Externalities Can Exacerbate Intergroup Inequality (Paul DiMaggio, and Filiz Garip, 2011)

實作程式碼：https://github.com/Letuvertia/Network_Externality

Overview

此篇論文探討*網路外部性* (Network Externalities, NE)與*同質偏好* (Homophily)是否會加速群體間的不平等。

我們實作了論文中提出的，基於美國調查的真實個體資料(家庭收入、種族與教育背景)建立的agent-based model去模擬採用網路(adopt internet)的群體行為，探討當個體的選擇會受到網路中其他個體影響時，以及不同的網路建立方式，整體上是否會加速群組間(分別依照家庭收入、種族與教育背景分群)的網路普及率(adoption rate)差距。

7種實驗條件

1 . No NE
2 . General NE
3~7 . Identity-specfic NE, **h= [0.0, 0.25, 0.5, 0.75, 1.0]**

Reservation price

依照論文Appendix，agent i的reservation price r_{it}定義為
$r_{it} = k \times y_{it}^{\gamma} + y_{it}^{\gamma} \times \delta \times n_{it}^{\alpha},\ k = 0.1,\ \gamma = \alpha = 0.5$，在實驗1中$\delta = 0.0$，在實驗2~7中$\delta = 0.1$，$y_{it}$為agent i的收入，n_{it}為目前網路中的adoption rate，其中網路可能為全域(探討general NE)或是為大小為agent i的network size的個人網路(探討identity-specific NE)。

Internet price

依照論文Appendix，網路價格為一個遞減到p_{min}的函數，定義為
$p_t - p_{t-1} = a \times n_{t-1} \times (p_{min} - p_{t-1}),\ a = 3.34/12,\ p_0 = 60.0,\ p_{min} = 28.74$。其中論文描述$n_{t-1}$為the number of subscribers in the previous period，但實作時我們發現如果設目前網路中的adopters數量，會導致函數大幅震盪，使模型完全不能跑。**我們認為這是筆誤，應該設為the percentage of adopters in the whole network in the previous period**，如此設可以使函數可以平穩下降至p_{min}，並可順利模擬。

Social networks and homophily

對於實驗3~7，每位個體會基於與他人的相似度建立個人網路，論文中將相似度定義為agent i和j在社經資料上的加權距離，formal form為
$sd(i,j) = || < W_I(Inc_i - Inc_j), W_E(Edu_i - Edu_j), W_R(Race_i - Race_j) > ||_2$。
$W_I = W_E = 0.53,\ W_R = 0.83$。在這五個實驗條件中分別有不同的同質偏好$h = 0.0, 0.25, 0.5, 0.75, 1.0$會影響網路的建立。建立agent i的個人網路的實作方式如下：讓i的network size為ns_i，首先計算i與所有其他的人之間的相似度，選出n個最接近的作為i的*in-group*，n設為ns_i的三倍，接下來i會建立ns_i次連結，他有$prob(T) = h + (1-h) \times prob_R(T)$的機率和*in-group*裡的人連結，否則和其他的人建立連結。實作上我們讓

$prob_R(T)$為$U(0,1)$，首先抽樣$prob_R(T)$計算$prob(T)$，接著從$U(0,1)$抽樣ns_i次並與$prob(T)$比較，來選擇與*in-group*的人連結的數量，最後從*in-group*中隨機抽取該數量並連結，其餘從其他的人隨機抽取。

Agent資料

論文。作者使用General Social Survey(GSS)在2002年的調查中2257位非裔美國人與白人受訪者的回答當作agent的社經資料，項目包含種族、教育背景、網路大小(network size)和收入。

實作。實作的資料來自於GSS Data Explorer。按照論文敘述，在網站上選擇了參數`race`作為種族，`educ`作為教育背景，`numprobs`作為網路大小，以及`income98`作為家庭收入。我們踢除了有任意一項為Not Answer或是N/A，或是種族為Others的資料，最後剩下2257位受訪者，與論文一致。

按照論文描述，在預處理上種族白人為1.0，非裔美國人為0.0;
教育程度不處理，單位為年；家庭收入由於回答是範圍，我們照均勻分布隨機分配給agent範圍內任意整數，對於回答Under 1000者設範圍為$[0, 999]$，對於回答110000 or over者設範圍為$[110000, 650000]$(650000的選擇解釋見論文註釋9)。最後將教育與收入normalized到$[0, 1]$，以和種族的範圍相符。實作上我們使用除以maximum的方式normalization。

比較。依照論文Appendix，種族與取log後的家庭收入相關係數為$.126$，種族與教育程度的相關係數為$.129$，教育程度與收入為$.290$。

我們以相同方法計算相關係數，其中log使用ln：種族與取ln的家庭收入相關係數為$.186$，種族與教育程度的相關係數為$.128$，教育程度與家庭收入的相關係數為$.283$。其中種族與家庭收入的相關係數差異甚大，由於收入的取值是範圍隨機取值，我們測試不同random seed，其相關係數皆在$.178 - .196$之間($.186$使用$seed = 2$)。

方法更動。在論文Appendix中提到reservation price公式中，k之所以選擇0.1是為了要讓在初始狀態沒有人有網路時(沒有網路外部性的影響)，只有1%的人能夠負擔得網路。依照論文提供的參數，網路初始價格$p_0 = 60.0$，反推只有家庭收入≥ 360000的人才負擔的起。在我們資料2257個agents中回答110000 or over的有214位，若按照論文參數從$[110000, 650000]$中抽樣，在期望值上會使得在初始時負擔得起的網路的人約有5%，**依據論文註釋9提到在他們資料中超過$110k$的約為10%，若$k = 0.1$則在期望上無法達到1%，敘述前後矛盾。為了符合1%的敘述，對於$> 110k$的人我們改成從$[110000, 385000]$抽樣。** 值得注意的是我們之所以不改動k而改抽樣上界，是因為論文提到δ需要設成和k一樣，目的是要讓income effect和network effect有相同的效果，如果調小k使得能負擔的人變少，則須同時調小δ，這會使得每個agent的reservation price上界變小，連帶影響internet price的參數設計，因此調整抽樣上屆僅會影響收入抽樣後為110k~360k的人的分布，而對其他人來說網路外部性仍有一樣的效果，是相對起來對模擬有較小影響的方式。

實作模擬流程

對於七種實驗條件，每一個實驗條件會模擬100個peroid。

準備階段

1. 建立 n=2257 個agents
2. 建立每一個agent的個人網路 (if 實驗3~7)
3. 初始化internet price和每一個agent的reservation price

Each Period t

1. 更新internet price
2. 每一個agent選擇是否要買網路
3. 更新每一個agent i的網路adoption rate n_{it}
4. 更新每一個agent的reservation price

Method of Evaluations

Odd Rates

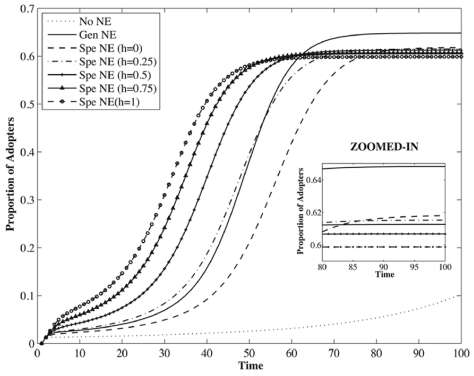
皆按照論文方式將三個變項各自分為兩群，Odd Rates的算法為group A的adoption rate/group b的adoption rate。

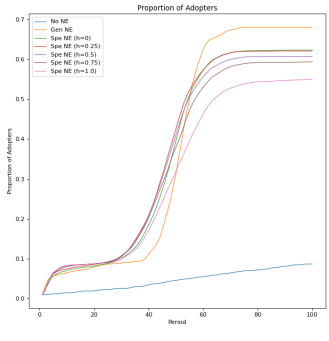
Logit Coefficient

論文中以logit coefficient去分析net effects，提到使用logistic regression在每一個period對agent-level adoption作prediction。在論文中沒有提到regression的X，但有提到三個變項各自對adoption rate的影響(各變項每上升一單位會對adoption rate造成多少變動)，因此在實作上，我們使用每一個agent的三個未normalized的變項，對收入取ln，作為X (size=(2257, 3))，每一個agent是否adopt作為y (size=(2257,))，在每一個period進行regression並紀錄coefficient。

實驗重製結果比較

7個實驗條件下的網路採用率 (adoption rate)





根據論文的觀察

1. general NE會達到最高的adoption level，因為每一個agent的adoption都會對其他每一個agent產生影響，導致他比起個人網路可以接觸到更多的agent。
2. 考慮個人網路(Specific NE)時，h越大會使他越快takeoff，但會達到較低的adoption level，因為h越大的網路agent個人網路的同質性會越高，使得一旦有人adopt，同溫層會快速跟著adopt，使得adoption rate快速上升，但也同時因為同溫層厚重，會較難接觸到其他的agent，導致最終的adoption level較低。

檢視重製結果，可同樣明顯發現觀察1；然而對於觀察2，我們發現確實h越大會使得adoption level越低，但並沒有發現h越大會有越早takeoff的情形。我們的解釋如下：要達到快速takeoff的條件是，在一開始的那群不需要靠網路外部性便可以負擔得起的人所擴散出的網路中，要打進那群密集連結的同溫層才能夠導致adoption rates的竄升，然而能夠滿足這樣的條件我們懷疑而非全然是因為h的升高，而是dependent on different dataset。換句話說，他們對於adoption level的觀察只是一個特例。

16. Beyond Social Contagion: Associative Diffusion and the Emergence of Cultural Variation (Goldberg, A., & Stein, S. K., 2018)

實作程式碼：https://github.com/anj226/Social_Contagion_Simulation

Overview

本篇論文的目的為探討 associative diffusion 對文化差異形成的影響。不同於傳統模型，本模型建構在沒有預先分群、沒有預設的社會網路的情況下，認為人與人之間傳遞的不是行為本身，而是對不同行為之間相容程度的觀感。

模型簡介

本模型的運作方式為每一輪隨機選取2個agents讓他們互動，因為建構在沒有預設網路結構的情況下，因此任兩人互動機率相等。其中一個agent A會展示兩個行為；另一個agent B觀察並受到展示者 A的影響，更新他對相關程度的看法 (associative perception) 並視情況調整偏好 (preferences)。若調整過後可使agent B對習慣相關程度的看法以及偏好的一致性增加 (相關程度高的偏好相近)，則保留該調整項，否則不更新偏好。論文裡先嘗試了只有兩個 agents 的模型，看兩者在互動後是否會產生相近或相反的偏好，再使用多個 agents 的模型進一步觀察兩兩之間的互相影響如何擴展到文化差異的形成。

模型架構

初始設定

1. 大小為 K 的習慣 (以下簡稱習慣) 集合。
2. N 個 agents，每個 agents 由兩個資料結構構成：

   a. 矩陣 R (大小K*K)： R[i][j] 代表他認為 習慣 i 跟習慣 j 多相關。初始值皆為1，假設agent一開始覺得任兩個相關程度都相同。

   b. 向量 V (K 維)：V[i] 代表他對習慣 i的偏好程度，初始值 ~U(-1, 1)。

迭代過程

1. 隨機選取兩agents A, B，由 A 作為表演者，B 作為觀察者
2. A 展示兩個習慣 i, j，各習慣被展示的機率由 V 決定 P(i) = $\frac{e^{V_i}}{\sum_{j=1}^{K} e^{V_j}}$
3. B 將 R[i][j] = R[i][j] + 1
4. B 將偏好較弱的習慣調整（較弱代表離平均值較近）得到 V'，調整的值 $\Delta v \sim N(0,1)$
5. 若 CS(V', R) > CS(V, R)，V' 保留，否則繼續使用 V
   - Constraint Satisfaction(CS) 的計算
     1. 將 V 轉換成一個$K \times K$的矩陣 Ω，其中 $\Omega_{ij} = |V_i - V_j|$
     2. 把 Ω 除以其最大值做標準化，同樣的將R矩陣除以其最大值標準化
     3. $CS(V,R) = \frac{K}{K(K-1)} \sum_{i=1}^{K} \sum_{j=1}^{K} |R_{ij} - \Omega_{ij}|$
6. R decays: R[i][j] = λR[i][j], where 0 < λ < 1 is the decay rate

模擬成果

衡量標準

Cognitive agreement

- interpretative agreement: 對相關程度看法的一致性
  - interpretative distance (interpretative distance下降，代表agents間的interpretative agreement上升)
- Evaluative agreement: 對行為偏好的一致性
  - Preference Similarity: 衡量agents 對行為偏好的相似程度
  - Preference Congruence: 衡量agents 對行為分類一致的程度

Behavioral agreement

- 使用 mutual information 來衡量行為的一致性 (計算方式有點複雜，如下)

  $I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$; 令 X = b1, Y = b2 計算兩者的 Mutual Information

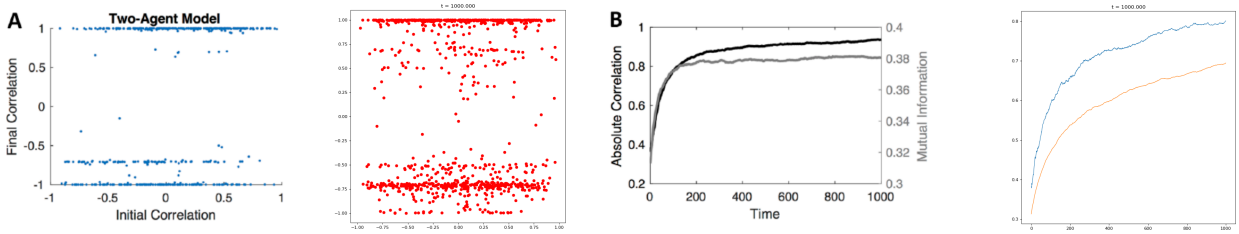
  其中 p(x) 為所有agent $P(b_1 = x)$ 的平均 $\frac{1}{N} \sum_{i \in N} P_i(b_1 = x)$，對agent i 而言$P_i(b_i = x)$ 即為根據偏好向量V計算所得的機率。同理， p(y) 為所有agent $P(b_2 = y)$ 的平均，但是因為模型有限定兩個行為不相同，因此$P_i(b_1 = x, b_2 = x) = 0$，且對agent i 而言，

  $P_i(b_2 = y) = \sum_{x \in K, x \neq y} P_i(b_1 = x, b_2 = y)$，其中 $P_i(b_1 = x, b_2 = y) = P_i(b_1 = x) P_i(b_2 = y | b_1 = x)$，而 $P_i(b_2 = y | b_1 = x) = \frac{P_i(y)}{1 - P_i(x)}$。

  整體而言，行為間的mutual information 如下：$I(b_1, b_2) = \sum_{x \in b_1} \sum_{y \in b_2} P(b_1 = x, b_2 = y) \log \frac{P(b_1 = x, b_2 = y)}{P(b_1 = x) P(b_2 = y)}$

Two-Agent Model

探討兩個agents互動的影響 --> 結果顯示兩者偏好會收斂或發散

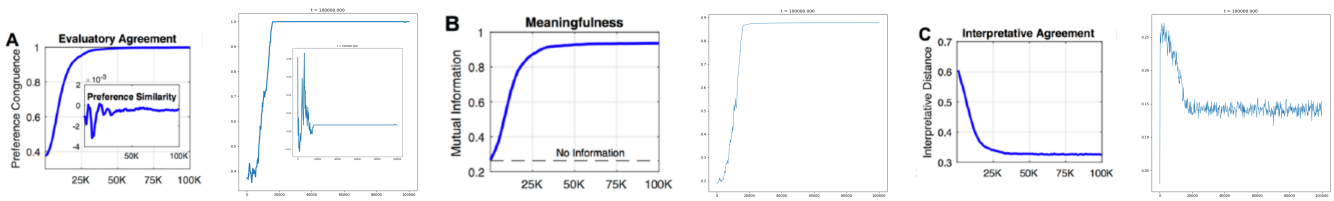


A. Final Pearson correlation between agents' preference vectors as a function of their initial correlation

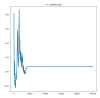
B. Absolute correlation between preference vectors and mutual information between the behaviors performed by each agent, as a function of time, averaged across all simulations

Multi-agent Model

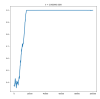
若兩個agents互動後會有相似或對立的偏好，預期多個agents會產生不同文化族群。



A. Evaluatory Agreement



- **Preference Similarity**
(measured as mean correlation between agents' preference vectors)
--> agents 有不同偏好



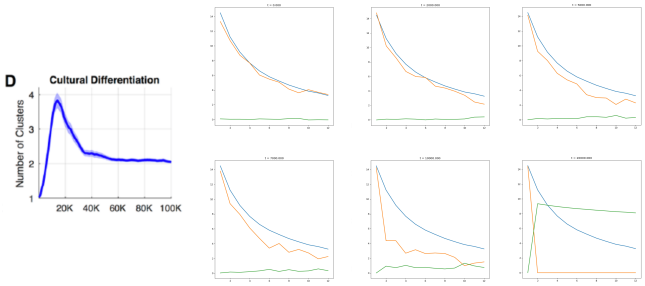
- **Mean preference congruence between agents**
(measured as absolute correlation between agents' preference vectors)
--> agents 慢慢分成兩個偏好相反的族群

B. Meaningfulness (Mutual information between agents' behaviors)
--> mutual information 上升，指示行為的意義上升

C. Interpretative Agreement (Mean distance between all agents' associative matrices)

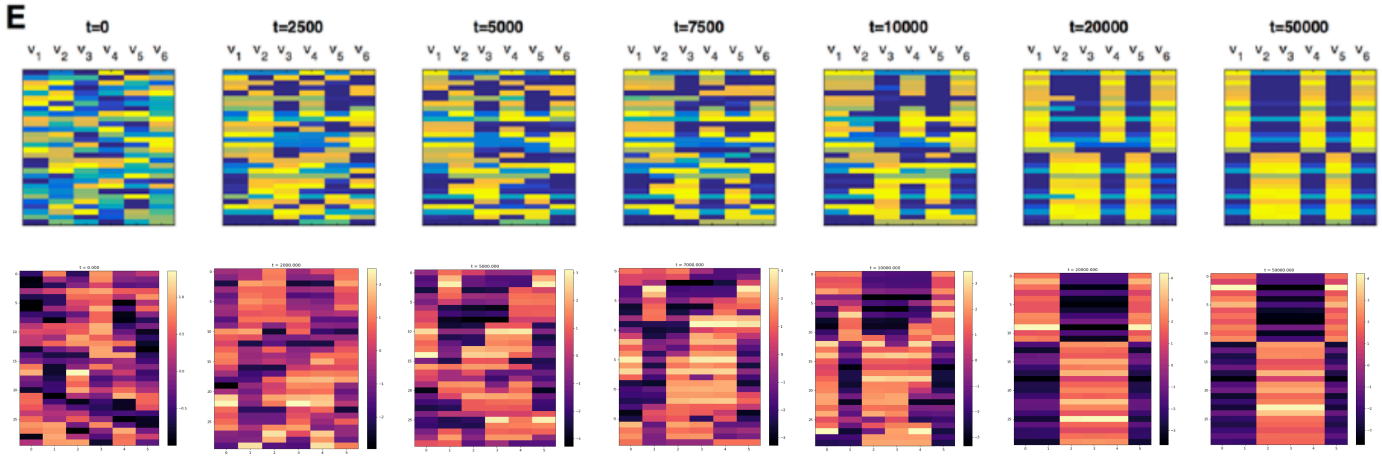
--> 對於行為關聯性的認知越來越接近
在圖形上是接近的（一開始是從亂數出發，所以t接近零的時候的資料基本上沒有意義），但如果注意座標軸其實會發現作者的圖比我們多了兩倍，我們猜只是公式計算上差了一個係數（由於作者也寫過 K/K(K-1) 等奇怪的係數，在加上其他的資料都蠻接近的，我們就先推測是少了一個係數。）

D. Cultural Differentiation



--> 對於分群的預測。因為找不到作者畫出連續函數的方法，我們只有用多次的 K-means 來感受作者所畫的圖的趨勢。t較小時，因為要測定的資料(黃線)、跟隨機出來的參考資料(藍線) 差別不大，所以 Gap值(綠線) 不明顯，到 t=10K 左右的時候，可以看到 Gap 開始有一些峰值，在差不多 4 個cluster 左右的地方，對應到作者圖片裡 t=10k 時那個主要的峰值。接著到 t=20k 之後，測定的資料(黃線)幾乎就變成兩堆了。(可以搭配 圖E 的變化)，我們的結果 似乎比作者更快趨近2個cluster(這點也可以在圖E看到)。

E. Snapshots of preference vectors (each heat map represents the preferences of 30 agents for 6 practices)



--> 可看出兩個族群的界線慢慢產生且逐漸分明 (由於是初始狀況是隨機分佈的，所以哪些習慣被分為一類並不是特別重要。重要的是上下兩個族群的分界，證明我們跟論文有類似的結果)

重製時遇到的問題，跟我們的應對方法：

1. 論文在其 **table 5** 裡 對執行 decay function 的敘述有點像是將 B 在該次迭代所觀察到的兩行為的關係乘上 decay rate，但我們認為不太合理。在此函式作用下，兩個完全沒有同時被觀察到的動作會永遠維持初始值 1，不會像預期中的趨近於零。依據論文內文的意思上判斷（也是我們認為較合理的做法），應該是不論有沒有被觀察到，整個 R 矩陣都應該隨著時間衰退，因此我們將 decay function 改成 R = λR。最後的結果也蠻符合他的圖，所以我們推測可能是表格裡對於 R 矩陣的下標的筆誤。
2. decay rate 在論文裡只有提及是一個介於零到一的數字，並沒有特定一個值。所以我們沒有辦法判斷作者所畫的圖是在哪一個 decay rate 底下所做的圖。所以我們嘗試了幾個不同的 decay rate 並發現確實會對結果的分析上有蠻大的影響的。從作者的圖我們覺得最接近的應該是 decate rate 為 0.8 的狀況，下面的圖也都是用 decay rate 0.8 所做的圖。
3. 對於 Constraint Satisfaction 的計算方式，我們覺得前面的係數有點困惑，我們的理解是他除以K(K-1)以求平均（扣掉i = j的部分），但分子的K 我們不太知道該如何解釋。論文中提到CS 是介於[0, 1]之間的值，而在標準化之後$|R_{ij} - \Omega_{ij}|$皆介於0~1，平均值亦介於[0, 1]之間。然而在乘以K以後，不能保證介於[0, 1]之間，因此我們以1代替其分母。但因為 CS 在此模型內僅用於大小比較，因此正係數實際上並不影響。
4. 論文裡 **figure 5** 的 **(D)**，透過把每一個 agents 的 V 向量（也就是對於所有行為的偏好程度）視為類似點座標的概念、用 1 - ρ(相關係數) 作為距離，他們能夠將 agents 分群，並透過 K-Means algorithm、 gap statistic method 這兩個算法，來計算時間對於 agents 分群數量。但因為K-Means algorithm，比較一般的用法是用在 N 維向量空間、距離定義為歐幾里得距離，但因為這邊座標為向量、距離定為(1-相關係數)，不太確定這樣子的情況下多個點的中心應該如何表達，這裡我們是採用多個向量求平均作為它們的中心。在gap statistic method 的部分，這個方法可以藉由多次的K-means以及用蒙地卡羅法抽樣出一個期望值比較，進而求出最優的cluster。但理論上應該是求出整數值，所以不太確定為什麼作者的作圖可以畫出連續的線。有思考過可能是多次計算後平均的結果，但在Gap statistic 裡面遇到了一些問題，有一些內容沒有寫得很清楚，作者在這裡附上了原論文，但因為時間因素，沒有辦法好好的去研究作者是怎麼畫出那條線的，所以就沒有做出這樣的圖了。但我們用多次的 K-means 所做出的圖，搭配另一個分群用的 elbow-method 可以看出作者的圖跟我們的結果是蠻符合的。
5. 本篇論文認為過往的傳統模型無法解釋文化差異的形成，因此在最後以一些圖表比較傳統模型（e.g. biased contagion）或者在已存在網路 (e.g. small-world network)下產生的分群狀況，但這些模型的的實作方法在本篇論文只有簡略介紹。例如在 biased contagion 的模型，提到其影響方式受到 bias parameter beta影響，但並無清楚介紹beta如何產生及作用，我們無法判斷他是每個 agent 固定且相同抑或是每一輪每個agent重新產生。而 small-world network 也只有解釋其大概模式，並無明確指示如何產生初始網路，因此難以實作。所以在時間考量上，我們就沒有再去實作那些部分。

結論

我們做出來的基本上在圖形上，都跟作者畫的圖形十分相似。但因為他從一開始初始到之後每一輪的決定agent、agent決定行為都有隨機性的關係，我們認為應該很難做出一模一樣的結果。在 two-agent的部分，因為是採用 1000次 模擬之後取平均，所以圖形上的表現會比較穩定，但仍然因為 decay rate 這個變數（論文裡沒有提到那些圖是在哪個decay rate下），會有一點差別。我們在寫 code 的時候，因為顧慮到這一點，我們的亂數是採用 python 裡面 numpy 模組的 random。並在開頭控制他的 random seed，這樣使得如果要重置一模一樣的結果，可以藉由同樣的random seed 再現。

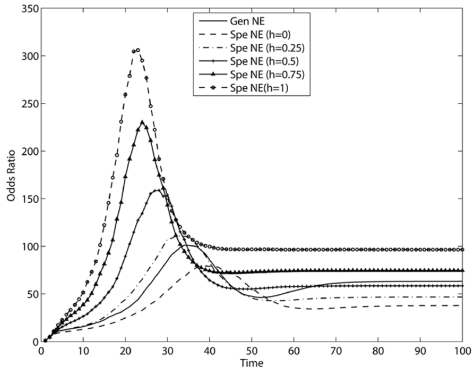
＊如果想要不能預測的結果，也可以將 seed 註解掉，但無論有沒有註解他都有好的亂數性質。

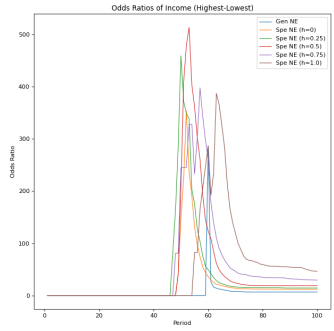
＊模擬以外，在觀測上可能也會用到亂數，就要避免用 numpy 模組裡的 random（可以換成 python 本身的 random）

附錄

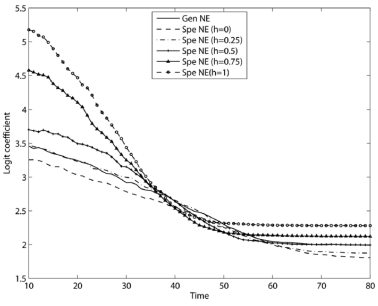
附錄一。論文9 的其他重製結果之圖表

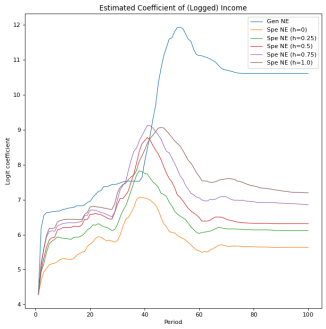
最高收入群組 v.s. 最低收入群組 - Odd Rates



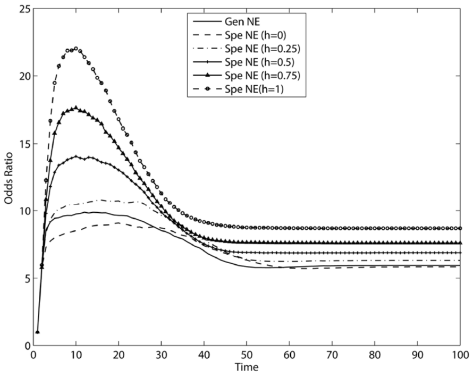


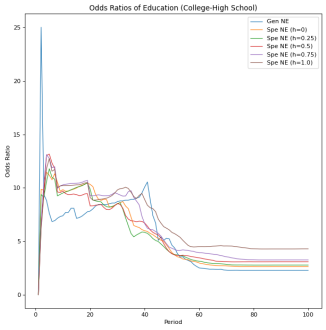
log(收入) - Logit Coefficient



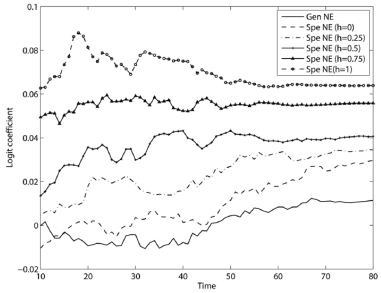


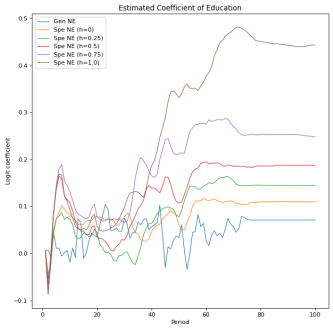
大學以上學歷 v.s. 沒有高中學歷 - Odd Rates



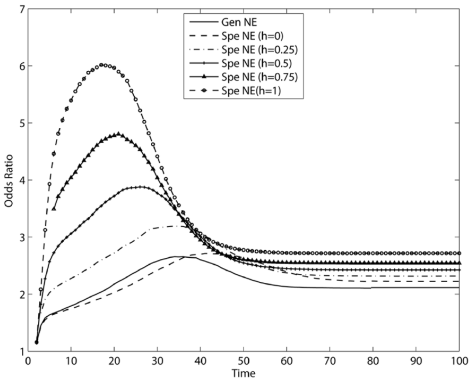


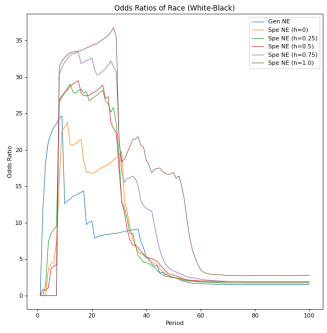
學歷年數 - Logit Coefficient





白人 v.s. 非裔美國人 - Odd Rates





種族 - Logit Coefficient