偵測網路檢舉系統中的共謀行為

中央研究院資訊科學研究所 / 數位典藏與數位學習國家型科技數位技術研發計劃

作者簡介

羅經凱

中央研究院資訊科學研究所多媒體網路與系統實驗室研究助理，研究方向為線上遊戲安全議題。E-mail: kaeaura@iis.sinica.edu.tw

陳寬達（陳存暘）

中央研究院資訊科學研究所助研究員及數位典藏與數位學習國家型科技數位技術研發計劃Web 2.0 團隊共同主持人，研究方向為網路量測、網路安全及線上遊戲。網址: http://www.iis.sinica.edu.tw/~cychen

摘要

目前網站審查違規行為的方法大多採取管理者以及網路檢舉系統，網路檢舉系統指的是當使用者發現惡意或者違規行為時，可對其提出檢舉，通知網站管理者處理。但是，檢舉系統可能遭到共謀者干擾，共謀者是一群為了私利或私仇合夥檢舉無辜者的用戶，這類不實檢舉使得沒有違規的使用者遭受檢舉，為當

事人或網站管理者帶來極大的困擾。在本文中，我們提出一套方法，以圖學演算法分析網路檢舉系統中的使用者相互檢舉關係，能夠偵測出真正的違規者以及意圖干擾系統的共謀者。

1. 引言

美國知名《時代》週刊評選出 2006 年度人物，不是具體的某個人，而是正在上網的「你」。當期週刊年度人物特刊的封面印著一台電腦，電腦顯示器螢幕上赫然寫著大大的“YOU”。《時代》週刊解釋說，現代社會正在進入“新數位時代的民主社會”，社會的重心正在由機構向個人轉移，因此 2006 年的當選者是正在使用網路或創造網路內容的每一個人。

Web2.0 網路社群的興盛，提供了表達自我與他人互動的新管道。擺脫了以往上網單向吸收資訊的情況，透過文字、照片以及影片，使用者得以將自己嶄露在世界的舞台上，也使網路上的人際交流更加熱絡。隨著使用者的網路創作以及人潮的湧進，許多網路社群網站上的資料量呈現爆炸性的成長，例如最近當紅的維基百科網站 Wikipedia、視訊共享網站 YouTube，以及微軟個人部落格網站 MySpace 這些新近湧現的網站。以 YouTube 為例，單日瀏覽次數便可達到一億次以上，這再再顯示目前 Web 2.0 的熱門程度。然而，每一件美好事物的背後都可能有它的缺點。Web 2.0 的基本精神在於「人際合作」，任何人都可以參與修改或者更新網站內容，藉由使用者的參與而展現出集體智慧的強大力

量。但是集體智慧的力量卻如同兩面刃，人們善意交流資訊的同時，有些人卻趁機散播謠言或者進行不軌行為。例如，在網路相簿中散佈違反善良風俗或者侵犯他人著作權或隱私權的照片、在視訊共享網站中提供血腥暴力的影片、在拍賣網站中販賣槍械或毒品、或在網路遊戲中使用外掛程式作弊等等。因此，網站內容合法性的維護是不可或缺的，除此之外，網路內容的維護也需要持續長時間的進行。目前 Web 2.0 網站常見的做法是僱用人員管理維護內容，然而對於 Web 2.0 網站如此龐大的內容與不斷的更新，勢必需要投入大量的人力以及時間才能夠維護網站內容。以知名社群相簿網站 Flickr 為例，此網站平均每小時收到網友上傳的照片 10,000 張以上，若單純以人工過濾每張照片是否違反網站規定或侵犯他人權利，以平均每小時一個人檢查100張照片計算，則Flickr至少需要僱用 1,000 位工作人員日以繼夜、毫不間斷地檢查，才能趕上使用者上傳的速度。

目前除了由網站管理者維護網站內容的合法性，另一種機制是開放使用者檢舉他們所發現的違規行為，回報通知網站管理者。如此一來，便能補足網路管理人力不足之憾，也省下大量的長期維護成本。然而，在大型的網站，如社群相簿網站Flickr，即便使用者檢舉的照片僅佔每日上傳照片數量的百分之一，那麼每天的檢舉回報量仍可達到 24,000 張以上的照片，這樣的數量對於網站管理者來說，依然是相當沉重的。為了不費人力審核，讓網路檢舉系統全程自動化，

我們希望能夠自動篩選出哪些被檢舉的違規者「惡行重大」，應該優先禁止其權限或移除不當內容。一個直覺的做法，我們稱作「多數票裁決法」，依照每位使用者的被檢舉次數進行排序，被檢舉次數越多，則認為其違規程度越嚴重。

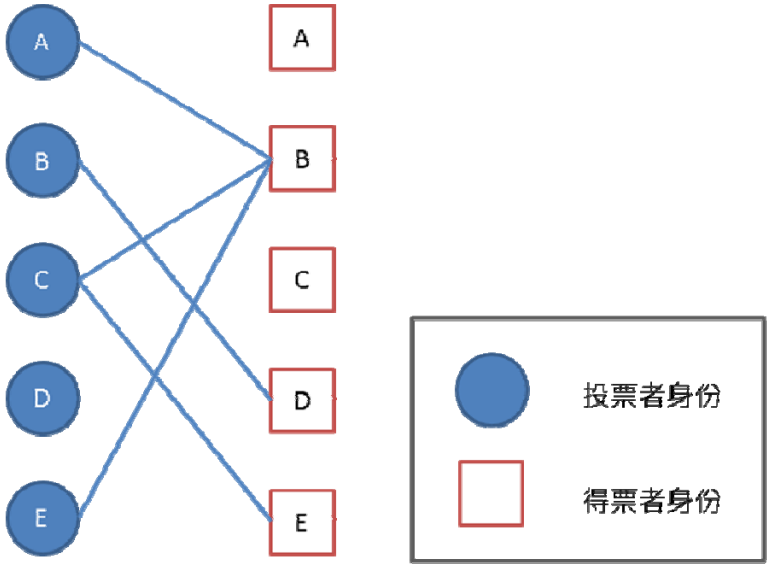
但是，單憑檢舉次數來決定違規者，潛藏遭到共謀者攻擊的隱憂。共謀者指的是一群勾結合謀的檢舉者，他們為了私利或私仇而去檢舉並無違規行為的無辜者。舉例來說，在網路遊戲中，某些玩家使用機器人程式賺取不當利益或者違反遊戲公平的原則，而遭到網友投訴檢舉後。被檢舉的玩家為了報一箭之仇，便私下找一群朋友，聯合投訴這些檢舉他們的玩家。這樣的共謀行為會造成兩種現象：其一，清白的使用者會遭到誣陷；另外，可能造成該被舉發的違規者，因其檢舉次數遠少於受到共謀者誣陷的使用者的舉發次數，而沒有被檢舉出來。

在本文中，我們提出一套方法，讓網路檢舉系統能在共謀行為存在的情況下，仍然能正確判斷哪些使用者為受到共謀者陷害的受陷者，哪些使用者為真正進行不軌行為的違規者。我們利用使用者之間的相互檢舉行為，找出誠實檢舉行為的特徵，再依循找出的誠實檢舉票顯現出真正的違規者。我們採用電腦模擬網站使用者之間的檢舉行為，藉此評估方法所檢驗並比較「多數票裁決法」方法與我們所提出方法的準確性。我們的方法與「多數票裁決法」皆有高達 90% 違

規者率判斷率外，在防治共謀者成功率的表現上，「多數票裁決法」僅能辨識出50%遭受共謀者陷害的用戶並還原其清白，而我們的方法則有高達 95%以上遭受陷害的用戶還原應有的清白。

2. 偵測違規者方法

檢舉系統內的使用者可以分為四種類型：由誠實檢舉者選出的「違規者」、共謀者所陷害的「受陷者」、遭到使用者誤判檢舉的「被誤判者」以及沒有違規行為也沒有遭受到檢舉的「良民」。我們將使用者的檢舉行為當成一種關係(relation)，並以圖(graph)表示，稱為「投票圖」。我們依據投票圖內各點分佈的疏密將投票圖分割成數組社群；這裡「社群」指的是投票圖中，點與點間稠密相連的子圖；相較於社群內部的邊，由社群內部連往外部點的邊則顯得較稀疏。若將社群與社群之間的邊稱之為「外部邊」，我們的分析指出，若一位得票者擁有較多「外部邊」連線，則其為違規者的機率較高。以下，我們將解釋此特質存在的原因，以及我們如何利用此特質來偵測違規者。

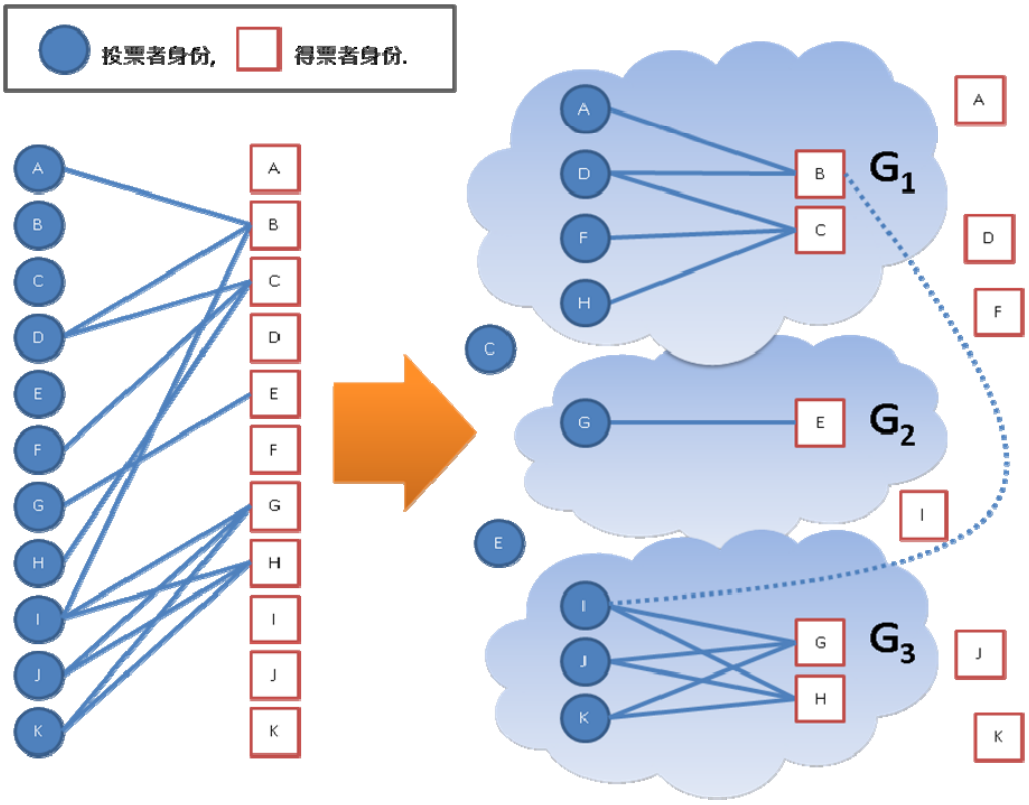


1 投票圖

我們定義「投票社群」為投票圖中點與點稠密相連的子圖，且該子圖連往其他子圖的邊會比子圖內部的邊稀疏。依此定義，同一社群的使用者有著緊密的相互檢舉關係，而不同社群的使用者之間的檢舉關係較為鬆散。若一群投票者在同一社群中，即表示他們擁有緊密的檢舉關係，同時也代表他們擁有相似的投票傾向。此指的「投票傾向」為在檢舉系統中使用者舉發的對象；若兩位使用者舉發的對象為同一群人，那麼我們稱這兩位檢舉者有相同的投票傾向；若兩位使用者舉發的對象中，有越多相同的人，則這兩位使用者有越相似的投票傾向。因此，每個社群皆由一群投票傾向相似的投票者與他們的檢舉對象所組成。

為了找出投票圖中的投票社群，我們採用 Newman 等人所提出偵測社群的演算法 Girvan-Newman Algorithm。此演算法的主要構想是在圖中找出不屬於社群核心的邊，藉由逐一移除這些邊，圖便被逐步地分割成多個社群。例如，在

圖 2 左方的投票圖可分割為三組投票社群。



2 左邊的投票圖經過分割之後，分出右邊三組投票社群，以及多個孤立點。

在投票圖中，若一條邊的投票者與被投票者分別落在不同的投票社群，我們稱這條邊為「外部邊」。外部邊擁有以下性質：

性質1、 該邊所表達的檢舉關係通常不是(共謀者, 受陷者)。

性質2、 該邊所表達的檢舉關係通常不是(誤投者, 被誤投者)。

性質3、 該邊所表達的檢舉關係極可能是(誠實檢舉者, 違規者)。

性質一：為了私仇或私利，共謀者會合夥檢舉受陷者，也因此通常擁有相近的投票傾向。根據社群的定義，投票邊稠密的使用者傾向形成同一個社群；因

此，共謀者與他們的檢舉對象很容易落在同一組投票社群中，所以社群與社群之間的檢舉關係邊通常不是(共謀者, 受陷者)。

性質二：即使是正常的檢舉者，也可能因無心或失誤而不小心檢舉到無辜的使用者。不論誤判機率為何，只要使用者數目夠多，可誤投的對象一多，一位使用者遭到多人誤投的機會將變得極小。因此，大部分的被誤投者最多只有一兩條連接邊，並與其誤投者落入同一個社群中。所以，外部邊是(誤投者, 被誤投者)的可能性並不高。

性質三：因為檢舉系統中只可能出現三種邊：(誠實檢舉者, 違規者)、(共謀者, 受陷者)以及(誤投者, 被誤投者)，基於性質一與性質二，我們可以推測出性質三：投票圖內大部份的外部邊是誠實投票邊(誠實檢舉者, 違規者)。

接下來，我們分析每位使用者的外部邊數目以及連接的身分(投票者或得票者)，以此來做為辨識違規者的依據。我們為每位使用者定義兩個特徵值：「外來票源數」以及「內部票源外出數」。

「外來票源數」指的是得票者獲得其他社群投票者的檢舉數量，也就是其所連接

的外部邊數量。我們的分析指出，當使用者的外來票源數量高的時候，表示其被較多的誠實投票者所檢舉，那麼他是違規者的嫌疑便越大。「內部票源外出數」指的是是社群內檢舉者投票給其他社群使用者的票數總量，我們定以此來量化得票者社群內部票源的誠實傾向。若使用者的社群外部邊的總數越高，該社群越可能是由一群誠實的投票者組成，其所檢舉的使用者便越可能是違規者。

根據以上分析，外來票源數或內部票源外出數量越高，使用者是違規者的可能性越大。我們以分群方法 K-means 根據此兩個值將所有使用者分為兩群，並判斷群聚中心距離原點較遠的一群為違規者。

3. 效能評估

據我們所知，目前相關研究中並沒有大量的投票資料供共謀行為偵測研究使用，因此我們採用電腦模擬來評估我們所提出的演算法。

模擬參數設定	
變數	**意義**
U	使用者人數
R	投票回合數
V	受陷者人數
P	一般使用者投票率
P_c	共謀者投票率
P_{error}	使用者的投票失誤率

T	違規者數量
C (C > 3)	共謀者數量

表 3

在我們的檢舉系統中，使用者投票身分種類包括誠實投票者與共謀者，而得票者身分包括受陷者、違規者及被誤投者。我們設定使用者的總數為 U，並隨機從 U 位使用者中選出 T 位為違規者；同時，我們也隨機設定超過三位以上 C 位使用者為共謀者，這些共謀者將共同陷害 V 位受陷者，其中 V<C。

系統內的投票規則為：

規則1、 誠實使用者檢舉違規者。

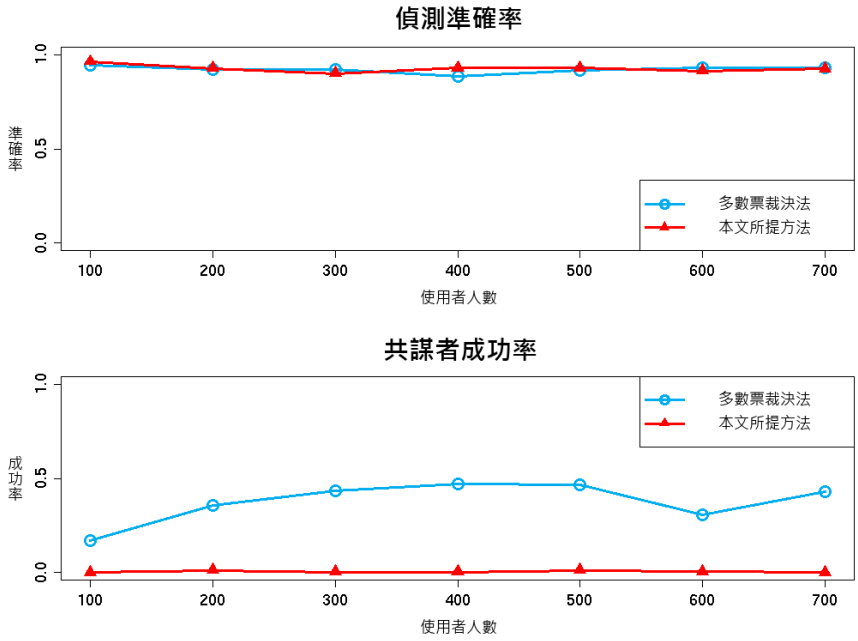
規則2、 共謀者檢舉受陷者。

規則3、 每一次的檢舉，都有機率 P_{error} 發生投錯票的情況。若發生此情況，誠實投票者將隨機檢舉非違規者，而共謀者將隨機檢舉非受陷者。

我們的檢舉系統模擬採取回合制，檢舉系統在運行 R 回合後結束，我們再套用違規者偵測演算法來辨識違規者及共謀者。在每個回合中，誠實使用者有機率 P 提出檢舉，共謀者有機率 P_c 提出檢舉。由於共謀者屬有意陷害受害者，所以主動檢舉的機率會比誠實使用者來得高，所以我們分別設定 $P = 10\%$, $P_c = 20\%$，而誤投機率 $P_{error} = 5\%$。最後我們將重複的檢舉票刪除，維持投票圖中每位投票者與得票者之間最多只有一條投票邊相連。我們將詳細的模擬參數列

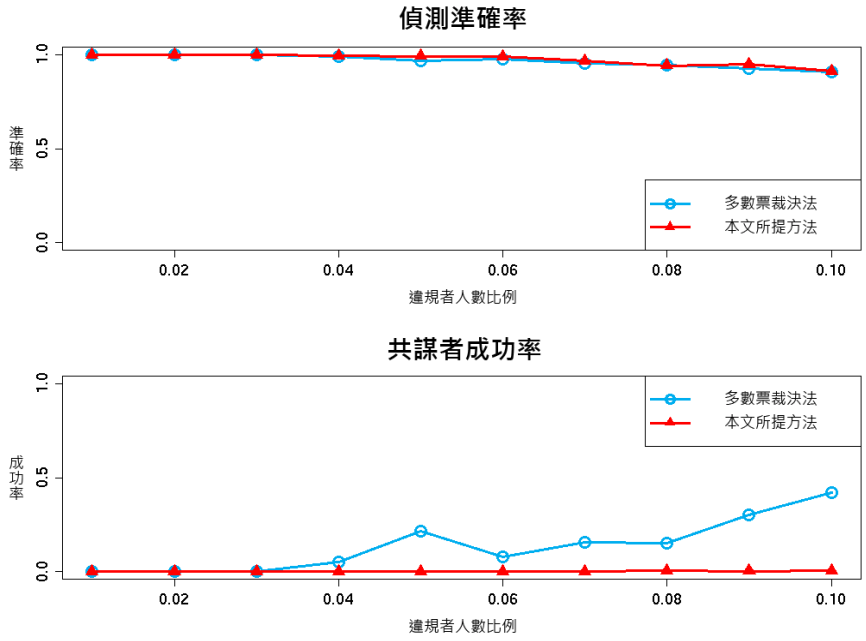
於表 3。

我們改變使用者人數 U 從 100 到 700 人，固定共謀者數量為 30 人，違規者數量為 40 人，並評估在這些情況下的違規者偵測準確率。圖 5 顯示，無論使用者數量為何，我們的方法與多數票裁決法的違規者偵測精準度皆穩定維持在 85% 以上。但是，在共謀者成功率的表現上，多數票裁決法顯得不盡理想，隨著系統人數增加，共謀者成功率最高可達到近 50%。若僅以得票數做為違規者的判斷依據，在受陷者與違規者得票數相近時，多數票裁決法將容易將受陷者誤判成受害者。而我們提出的方法以使用者得到的誠實檢舉票數為依據，所以即使受陷者得到不少共謀票，受陷者也不容易被誤判成違規者。結果顯示，即使使用者數目增加到 700 人，根據我們的方法，共謀者成功率不曾超過 5%。



圖表 4

另一方面，為分析違規者比例的影響，我們將模擬人數設定為 400 人，共謀者數量30人，違規者人數由4人(佔所有使用者人數1%)逐漸增加至40人(10%)，並計算違規者偵測方法的正確率。如圖 6 所示，隨著違規者數量比例由 1%上升到 10%，兩個方法的精準度都有略微下降的趨勢，不過皆維持九成以上的正確率。多數票裁決法在違規者比例超過 4%之前，保持著近乎 0%的共謀者成功率。主要的原因是，當違規者數量少時，誠實檢舉者的票容易集中落在少數違規者身上。如此一來，違規者的得票數通常會遠遠超過受陷者的得票數，因此多數票裁決法不會誤判受陷者為違規者。但若違規者比例升高，每位違規者得到的票數相對變少，單以得票數判斷違規者的方法將會容易產生誤判(受陷者為違規者)。由圖 6 可看出，在違規者人數超過 4%的情況下，多數票裁決法的共謀者成功率開始上升，表示其容易將受陷者誤判誠違規者。另一方面，就算違規者人數比例上升到 10%，我們的方法依然可保持共謀者成功率低於 5%。



圖表 5

4. 結論

在本文中，我們討論網路檢舉系統在 Web 2.0 網站的必要性，以及為了降低人工審核需求，衍生自動化審核檢舉案件的動機。在檢舉系統中，共謀者為了惡作劇、私利或私仇可能檢舉清白的使用者，作出不實檢舉，這些檢舉使得我們無法單純地判斷高得票數的使用者即為違規者。為了不使共謀行為混淆檢舉結果，我們必須偵測出檢舉系統中的共謀行為，分辨出真正的違規者以及清白的受害者，同時抓到共謀者，讓檢舉系統達到原本供使用者檢舉不軌的目的。我們以電腦模擬評估我們的方法的偵測效能。結果指出，無論使用者數量如何變動、或共謀者或違規者比例如何提升，我們的方法皆可使共謀者成功率壓制在5%以下。反之，單純以用戶檢舉票的「多數票裁決法」法，會因為受陷者共謀票的增加而無法精準的判斷受陷者或違規者，使得共謀者成功率最高可達到近

50%。

大型的社群網站及論壇中，由於使用者上傳的文字、照片、影音等數量過於龐大，單憑管理者來審核上傳內容是不夠的；因此，這些網站通常提供會員檢舉功能，如果使用者看到違反善良風俗或違反版權、著作法的內容時，就可即時檢舉回報給管理者。例如，在 Yahoo 奇摩交友中，每個帳號主頁的右上角都有一個「檢舉申訴」按鈕，只要按下它，再填寫理由，就會由管理來重新審查被檢舉的使用者寫撰寫的內容或照片。這是一個好的設計，讓每一個人的看法都可以直接回饋到管理端，但另一方面，錯誤的 / 誤判的 / 過於主觀的，甚至是蓄意謀害的檢舉回報也履見不鮮，這使得管理者們必須花更多的時間精力來審核可信度並不高的檢舉要求，同時也可能因為工作量的增加降低審核的品質。我們在本文中提出的共謀行為偵測演算法即是為解決此問題而存在，一旦有精準的共謀行為偵測演算法，我們便可將檢舉系統「自動化」，自動列出犯行重大的被檢舉者，同時可自動忽略或禁止共謀者的投票行為，讓真正的違規者能在最快時間藉由誠實網友的檢舉被發現，達到網路檢舉原有的目的。