

DATACLYSM

无人旁观时 我们是谁

线上数据与线下行为
如何重构个人身份认同

Who We Are
(When We Think No One's Looking)

大数据下的
人类真实面目

[美] 克里斯蒂安·鲁德尔 _ 著
(Christian Rudder)

蒋宗强 _ 译

中信出版集团

無人旁觀時我們是誰：大數據下的人類真實面目

Dataclysm: Love, Sex, Race, and Identity--What Our Online Lives Tell Us about Our Offline Selves

出版社：中信出版社

作者：(美)克里斯蒂安·魯德爾

譯者：蔣宗強

出版日：2020/12/01

ISBN13：9787521720730

為什麼臉譜網上的點贊情況可以預測一個人的性取向，

甚至可以預測一個人的智商？

為什麼容貌姣好的女子獲得的面試機會比普通人多？

為什麼有些人對你的討厭反而促使其他人更喜歡你？

內容簡介

那些我們意識不到的數據信息滾滾而來，告訴我們是如何奮鬥、如何戀愛、如何變老的，以及我們是誰，我們正在如何變化。

大數據時代，即使無人觀察，但哪怕一個微小的網絡動作也能揭示出我們的行為方式。

數百年來，我們只能依靠民意調查或者小型實驗去研究人類社交行為。今天，一種新的研究方法成為了現實，那就是大數據觀察。

隨著我們的生活越來越多地依賴網絡，我們終於可以直觀地瞭解自己的生活，通過分析社交網絡上的大數據，作者向我們展示了人們在公開場合以及私下場合的自我表達方式以及社交方式。

本書為我們審視自我提供了一條新路徑，讓數學具有了人性化特徵，並演繹出了我們這個時代的方方面面。

作者簡介



克里斯蒂安·魯德爾 Christian Rudder，OkCupid交友網站的聯合創始人兼總裁，同時也是備受歡迎的博客OkTrends的作者。畢業於哈佛大學數學系，後來擔任SparkNotes的創意總監。其作品在《紐約時報》《紐約客》等媒體上反響熱烈。

無人旁觀時我們是誰：大數據下的人類真實面目

【美】克里斯蒂安·魯德爾 著 蔣宗強 譯
中信出版集團

目錄

前言

第一部分 我們因何而聚

第一章 伍德森法則

第二章 出醜效應

第三章 「作家」的世界

第四章 社交圖譜

第五章 「約會大冒險」：雖敗猶榮

第二部分 我們的隔閡從何而來

第六章 混淆變量

第七章 被神化的美貌

第八章 隱祕的選擇

第九章 憤怒的時代

第三部分 影響身份認同的因素

第十章 你是誰？

第十一章 你墜入愛河了嗎？

第十二章 瞭解自己所處的位置

第十三章 個人品牌

第十四章 蛛絲馬跡

後記

關於本書數據的說明

[致謝](#)

[版權信息](#)

前言

到目前為止，你肯定聽過很多關於大數據的事情，比如，大數據潛力巨大，會引發不利的後果，會顛覆很多領域的固有模式，併為人類開創諸多新模式，人們一直非常熱衷於關注大數據的網站，等等。提起大數據，你可能感覺頭暈目眩，就像頭部被什麼撞擊了一下似的，所以，我在這裡不炒作或報道大數據引發的各種現象或後果，但我闡述的所有內容都是圍繞著大數據本身展開的。在本書中，我為讀者呈現了自己蒐集的大量真實數據。我之所以能夠獲取並分析這些數據，是因為我交了好運，付出了大量勞動，並努力說服用戶，其中運氣的成分是比較大的。

我是交友網站OkCupid的創始人之一。經過我們長達10年的艱辛努力，它已經成為世界上最大的交友網站之一。該網站是我和三個朋友一起創辦的。我們都是數學系畢業的。這個網站之所以能成功，在很大程度上是因為我們能運用數學思維來解決約會問題。在歷史上，愛情問題往往是一些所謂情感「專家」的專屬領域，而我們卻能運用數學算法對愛情問題開展嚴謹的分析。我們這個網站的工作原理並不複雜，只需要通過一些嚴謹的算法來模擬兩個人相互瞭解的過程就足夠了。我們的方法引發了廣泛的共鳴，2014年，就有1 000萬人通過這個網站去結交朋友。^[1]

我非常清楚，很多網站（和網站的創始人）為了說明自己多麼受歡迎，往往喜歡拋出一大堆數據，而毫無疑問的是，大多數有思想的人已經逐漸學會了忽略這些數據。他們所說的數百萬這、數十億那，帶著一連串零的數字令人眼花繚亂，但這些基本上是為了往自己臉上貼金而自吹自擂。與谷歌、Facebook（臉譜網）、Twitter（推特）和其他網站相比，OkCupid遠遠算不上家喻戶曉的網站，如果你和你的朋友們已經結婚很多年了，而且婚後一直過著幸福的生活，那麼你們可能從來沒有聽說過我們這個網站。很多人可能從來沒有用過這個網站，而且不屑於參與一些初創網站的用戶體驗調查，因此，關於如何清楚地為這些人描述我們這個網站，我著實費了一番苦心。我會通過非常通俗易懂、富有感情的方式來描述。現在，每天大約3萬對男女在OkCupid網站的幫助下進

行第一次約會；^[2]大約3 000對男女在約會後會建立長期的戀愛關係；之後，其中200對情侶會走進婚姻的殿堂，而且很多人會生寶寶；今天，有些夫妻的孩子已經長大了，被父母逼著穿鞋的時候都會噘起小嘴表達不滿了。如果沒有我們的網站，這些小傢伙兒可能不會出生。

我絲毫沒有自以為是地認為我們一切都做得非常完美。可以說，雖然我和朋友們創辦的網站讓我感到非常自豪，但說實話，我並不在意你是不是我們的會員，是不是要在我們的網站註冊一個賬戶，我本人以及其他創始人都沒有過在線約會的經歷。如果我們的網站不適合你，我也能理解。請相信這一點。我最不喜歡做的事情之一就是像一個傳教士那樣向他人宣揚技術帶來的福音，我在這裡也不會秀出一連串令人眼花繚亂的數據來搶奪他人的寶貴地盤。我現在仍然在訂閱報紙雜誌，比如《紐約時報》週末版，在Twitter上發佈信息時，我會感到尷尬和侷促不安。我不會勸你增加或減少對互聯網或社交媒體的使用、尊重或信任。你完全可以堅持你對網絡世界的固有看法。但通過這本書的描述，我真心希望能說服你做一件事：反思自我。這就是我寫這本書的真正目的。OkCupid網站只是我實現這個目的的一個方式。

自2009年以來，我一直領導著OkCupid網站的數據分析團隊，我的工作內容是分析我們的用戶創建的數據。創辦網站的所有工作幾乎都是我那三個合夥人完成的，我這些年來只是擺弄林林總總的數據。我所做的工作中，有些有助於我們經營網站，比如，瞭解男性和女性對性與美的不同看法，對於交友網站的運營具有至關重要的作用。但我的許多分析結果只是有趣，並沒有什麼直接的用途。很多時候，我們雖然能夠通過分析統計數據發現一些事實，但我們卻無法採取什麼措施來改變這些事實。比如，通過數據分析，我發現蘇格蘭著名獨立搖滾樂隊——貝爾和塞巴斯蒂安樂隊（Belle & Sebastian）是世界上最陽光的樂隊。再比如，在晚宴上拍照時，使用快照功能會讓人看起來比實際年齡老7歲，除非讓嘴做出「哈」的形狀。我們在數據分析過程中發現的現象基本上都屬於這一類，雖然比較有趣，卻無直接用途。我們偶爾會把分析結論發佈出去，但這些結論似乎顯得無足輕重，也沒有引起大量關注。然而，當我們分析了足夠多的數據後，一些大趨勢就會變得明朗起來，就像很多小圖案拼接起來就會變成一個明顯的大圖案。我發現這個工作一個比較好的地方就在於可以直接通過觀察來分析一些禁忌性問題，比如種族問題。也就是說，我不必按照社會科學領域內的傳統做法去請求人們回答某些設計好的問題，也不用設計什麼小型實驗，而是直接分析用戶在網站上創建的數據就可以了。通過這些數據，我就能看到現實中發

生的事情。比如，我可以通過我的交友網站來觀察10萬名白人男子和10萬名黑人女子的私人交往情況。這些數據就在我們的服務器裡，這是一個令人無法抗拒的良好機遇。

隨著我持續不斷地進行數據挖掘和分析工作，發現也越來越多。於是，我開了一個博客，名為OkTrends，與世界分享我的發現。我所知道的永遠比我分享的多。後來，我整理了一下博客中的內容，並在此基礎上做了重要改進，於是本書便出爐了。為了寫這本書，我引用的數據源遠不止於OkCupid網站的數據。事實上，當代大部分重要的在線數據源都成了我的搜索目標，而且我掌握的關於人際交往的數據更加深入廣泛。在這本書中，豐富翔實的數據不僅披露了網站用戶的習慣，還揭示了一系列具有普遍性的行為模式。

人們在公開場合討論大數據時，主要集中在兩個方面：一個是政府利用大數據技術搞偵察監控活動，另一個是大數據技術對於商業機遇的影響。關於第一個方面，我並不認為我比你知道的情況多，我跟你一樣也只能通過看新聞了解冰山一角。據我所知，國家安全機構從來沒有為了使用某個交友網站的數據而專門同網站接洽。對於這些信息，國家安全部門不會很感興趣。不過，關於第二個方面，也就是如何利用大數據來賺錢，我瞭解得比較清楚。當我剛開始寫這本書時，各大網站和報刊的科技新聞板塊充斥著Facebook要上市的信息。Facebook收集了所有用戶的個人數據，利用這些數據發了大財，而今他們又要上市圈錢。2012年5月18日，Facebook上市，但《紐約時報》在5月15日的一條新聞說明了一切，該新聞的標題是「Facebook必須將數據轉化為盈利」。你或許能預料到會有一群人登上各大媒體的評論版面，發表一些評論文章，告訴美國人買Facebook的股票是一樁穩賺不賠的買賣。

我們OkCupid網站是依靠廣告收入支撐的。作為這個網站的創始人，我可以肯定地說，數據有利於提高銷售業績。網站的每個頁面都能捕捉到用戶的體驗。用戶點擊的每一個位置、輸入的每一條信息，甚至是一個頁面上停留的時間，都是用戶體驗的表現。根據這些，不難看清用戶的喜好以及如何滿足他們。這簡直太棒了！但當我看到某個用戶為其朋友們提供關於身體噴霧的最新消息時，我不會利用這個機會向他推銷身體噴霧。雖然我能接觸到這些用戶創建的數據，知道用戶喜歡點擊什麼、輸入什麼以及在某個頁面上停留多久，但我不會利用這些數據去做生意。長期以來，大數據主要服務於監控與賺錢這兩個目標，而在過去的三年裡，我卻努力利用大數據實現第三個目標，即反思人類的故事。

Facebook可能知道你是M&M巧克力的粉絲，然後給你推送與這種巧克力有關的促銷信息。如果你和男朋友分手後搬到了得克薩斯州，在Facebook上分享了很多與前男友的合照，並開始了新的約會，那麼Facebook就能掌握得一清二楚。如果你在谷歌上輸入一些尋找汽車的信息，谷歌會根據你的搜索記錄揣測你的心理，為你推薦一些品牌和車型。比如，如果谷歌知道你是一位傾向於追求刺激、B型血、25~34歲的男性，那麼谷歌可能會自動為你推薦斯巴魯汽車。

與此同時，谷歌也知道你是否是同性戀、是否生氣、是否孤獨、是否存在種族主義思想或者是否正在為母親的癌症而憂心如焚。Twitter、Reddit（熱迪網）、Tumblr（湯博樂）、Instagram（一款圖片分享應用程序）等首先屬於企業，具有企業的屬性，但與此同時，它們具有「人口統計學家」的屬性，而且它們統計的廣度、深度及重要程度都是史無前例的。現在，大數據能夠呈現出我們是如何奮鬥、如何戀愛、如何變老、我們是誰以及我們正在如何變化的。對於這些，人們幾乎是意識不到的。我們需要做的只是觀察。在大數據時代，雖然我們覺得沒有人觀察自己，但哪怕一個微小的動作也能揭示出我們的行為方式。下面我將詳細講一講我都觀察到了什麼。再重複一遍，我絕不會利用大數據搞推銷。

∞

本書講述的內容雖然與人有關，但僅僅從宏觀視角出發，概括性地分析了規模龐大的數據，幾乎沒有具體提到任何人的名字。本書運用了大量的圖表，也幾乎沒有涉及人名。現在的大眾科學在解釋問題時存在「以小見大」的傾向，即利用一些微小而古怪的事物作為透鏡來闡釋大事件，結果充斥著陳詞濫調。比如，認為一個蘿蔔折射了世界歷史，一條魚引發了一場戰爭，一個手電筒照射稜鏡便能讓臥室的牆壁上呈現出美麗的彩虹，等等。我闡述問題的方向與此相反，或者說是「以大見小」的。我擁有的大數據以萬億字節計算，這些數據都是關於人們行為、思想和言語的，是過濾了很多小事情之後得出來的。這些事情包括你的朋友圈談及你的婚姻是否穩定的言論，亞洲人（以及白人、黑人和拉丁美洲人）最不願意用哪種方式來描述自己，同性戀者都在哪些地方祕密聚會，人們的寫作習慣在過去10年發生了哪些變化。我們這樣做的目標就是促使人們在瞭解自我的過程中，減少對敘事方式的依賴，更加重視數字的重要性，或者說，我們要形成「數字也是敘事方式」的思維方式。

這種方法是從漫長而艱苦的統計工作中總結出來的。這本書凝聚了我與合夥人多年的努力。一個交友網站必須讓不同的人會聚到一起。要做到這一點，必須瞭解人們存在哪些慾望、哪些習慣以及厭惡什麼。所以，你必須蒐集海量的詳細數據，並努力將其轉化為具有普遍適用性的人類行為理論。每天同這些紛繁複雜的數據打交道，絕對不同於籌備一場婚禮或者編輯某個報紙的版面。在這些數據中，我瞭解到的是整個人類的一般情況，而不是一兩個人的具體情況。你瞭解人類之後，對人類的愛便油然而生。

因此，所有網站，也就是所有數據科學家，必須用計算機能夠讀懂的语言來客觀地描述人類行為。然而，對於非數字類的事情，用數學算法來處理並不是非常奏效。因此，如果你想讓計算機明白一個想法，必須盡己所能地將其轉化為數位。網站和應用程序面臨的挑戰就是如何分割連續性的人類行為，將其分割為一個個小片段，再收集起來裝進一個個小桶裡面，同時又不讓別人注意到這個過程。也就是說，Facebook、Reddit等網絡社區將人類的友誼、愛情分割成服務器能夠理解的片段。在數字化的同時，他們還要儘可能地把網頁界面做得貼近現實，讓用戶覺得你所提供的信息代表了真實生活。互聯網會給人造成一種微妙的錯覺。你可以想象一下，一個胡蘿蔔被切成一截一截的，整整齊齊地擺放在砧板上，看起來仍然像一個完整的胡蘿蔔，其實這只是一種錯覺。人類行為的連續性與數據庫的間斷性之間的矛盾是網站運行過程中面臨的一個複雜挑戰，而我講述的內容恰恰就是關於這個方面的故事。新技術的誕生為我們通過數字化方式分析人類慾望和友誼提供了一個新機遇。我們可以利用一些確鑿的數據來分析持久存在的難解之謎，也可以分析人類之前認為無法量化的活動，從而對這些活動獲得一定程度的瞭解。技術發展得越來越好，對人類生活的影響也日益廣泛，人類對技術的理解呈現出了令人驚訝的提升態勢。我在後文會舉例說明這一點，但我必須首先說明一點：我們OkCupid網站真的不打算把「分析無法分析之事」作為自己的宣傳噱頭。

在互聯網上，排名隨處可見。Reddit為用戶提供了「頂」或「踩」的選項，亞馬遜網站可以發表客戶評論，甚至Facebook也為用戶提供了點「贊」選項。這些網站之所以讓用戶投出自己的一票，表達自己的看法，是因為它們將這些動態的、個性化的事物轉化為它們能夠理解和利用的事物。交友網站之所以讓人們互評，就是因為一旦人們表達了自己對他人的第一印象，比如「他的眼睛很漂亮」，「哦，他很可愛，但我不喜歡他的紅色頭髮」，「噢，太難看了」，等等，網站就會按照滿分

為5分將這些評價轉變為一些簡單的數字，比如5、3、1等。網站收集了數足夠多的評價之後，就瞭解了一個人給他人留下的第一印象是什麼。如果將所有個人評論綜合到一起分析，就能以小見大，揭示出一些大趨勢，從而清楚地看到人們如何做出對他人的評價意見。

對於這些人與人之間的相互評價，你需要做的最基本的事情就是計數，也就是統計一下多少人得了1分、多少人得了2分、多少人得了3分等，以此類推，然後對你的統計結果進行對比分析。我曾經統計了男性對女性的坦率評價（這些人都不是同性戀者）。根據統計結果，我繪製了這幅柱狀圖（見圖0—1）：

這個簡單的圖形是根據OkCupid的5 100萬名用戶的偏好繪製出來的。從本質上講，它體現了男性對女性美貌的偏好，糅合了所有的小故事（即成千上萬名男性對女性的看法）與所有的逸聞趣事，最後形成了一個直接明瞭的圖形。通過這種全新的方式來看個體，就像從太空看地球上的人一樣，不會看到細節，但你會從中看到一些自己覺得熟悉的東西。

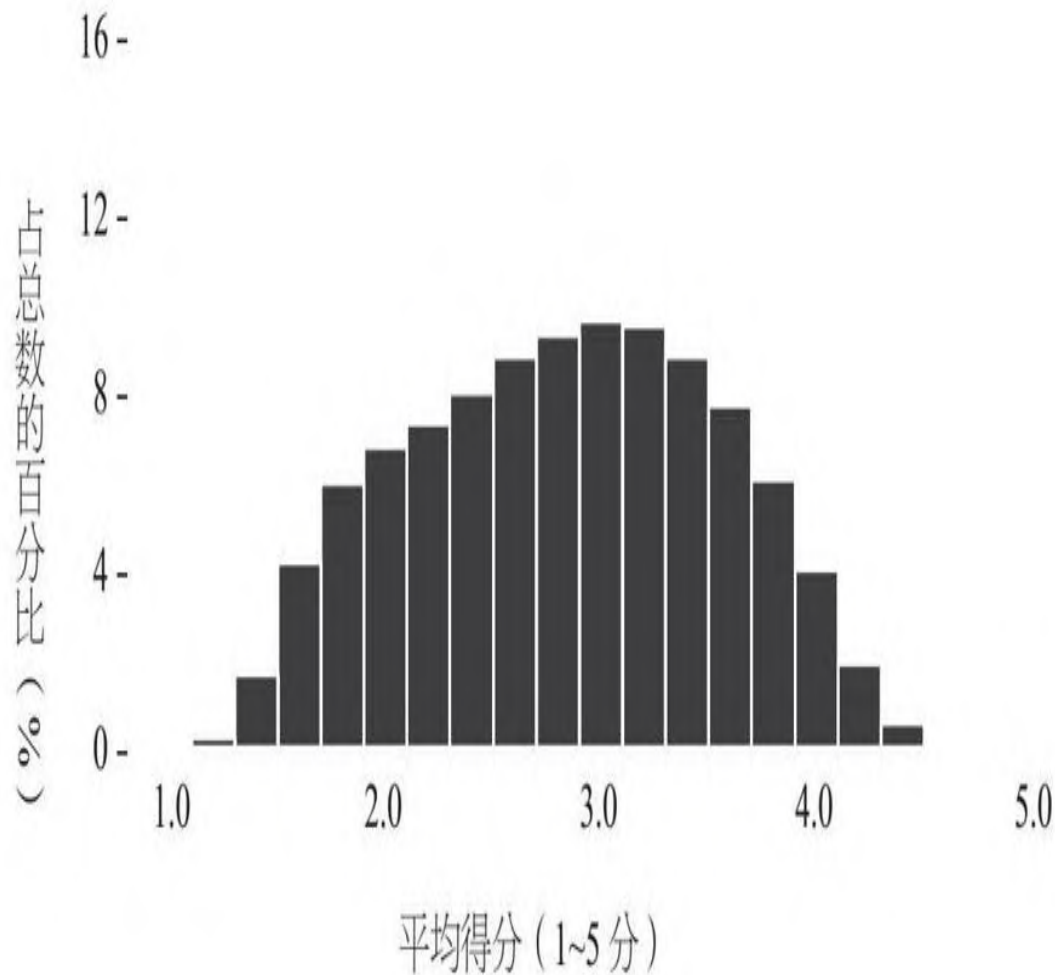


圖0—1 被男性評價的女性

這個圖形為我們揭示了什麼？我們很容易認為得到這個基本的鐘形曲線（柱狀圖形成的曲線）是理所應當的，因為很多教科書裡的例子可能會使你預料到這一結果。但在很多情況下，用戶評價不會符合人們的預期，尤其是當涉及個人偏好時，實際評價可能不符合我們的預期，可能呈現出「一邊倒」的情形，即評價集中在高分區。比如，Foursquare（一家基於用戶地理位置信息的手機應用服務網站）用戶對紐約比薩店的評價往往是非常積極的（見圖0—2）。^[3]

我們再看一下美國國會在主流媒體民調中的支持率。^[4]從道德角度來看，政客處於比薩店主的對立面，所以，政客的支持率就呈現出了另外一番情景，即評分集中到了低分區（見圖0—3）：

我們網站的男性用戶對女性用戶的評分呈現出了單峰曲線的態勢，這意味著女性的得分往往集中於單一值。這也是很容易就能預測到的，我們可能不以為然，但是很多情況下會出現多重模式，或者說有多個「典型值」。如果你根據NBA（美國職業籃球聯賽）球員在2012—2013賽季中出現在首發陣容中的頻率對其進行描述的話，那麼你會發現少數球員會頻繁出現在首發陣容中，而有些球員則幾乎沒有或從來沒有出現在首發陣容中。如果根據這個頻率製作一幅柱狀圖，就會發現少數球員會集中在一端，而中間部分幾乎沒有人（見圖0—4）。[\[5\]](#)

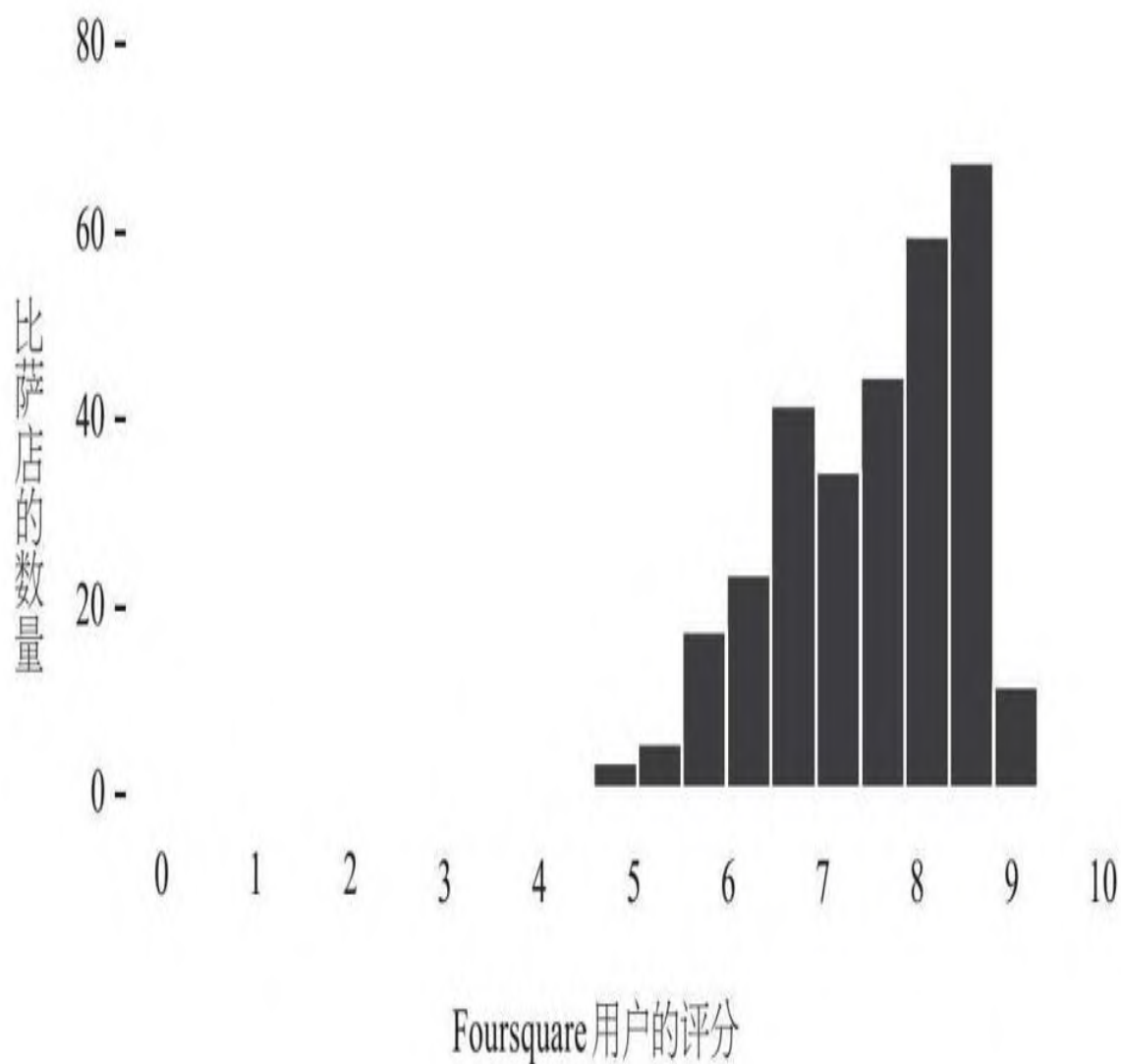


圖0—2 用戶對紐約比薩店的評分（1~10分）

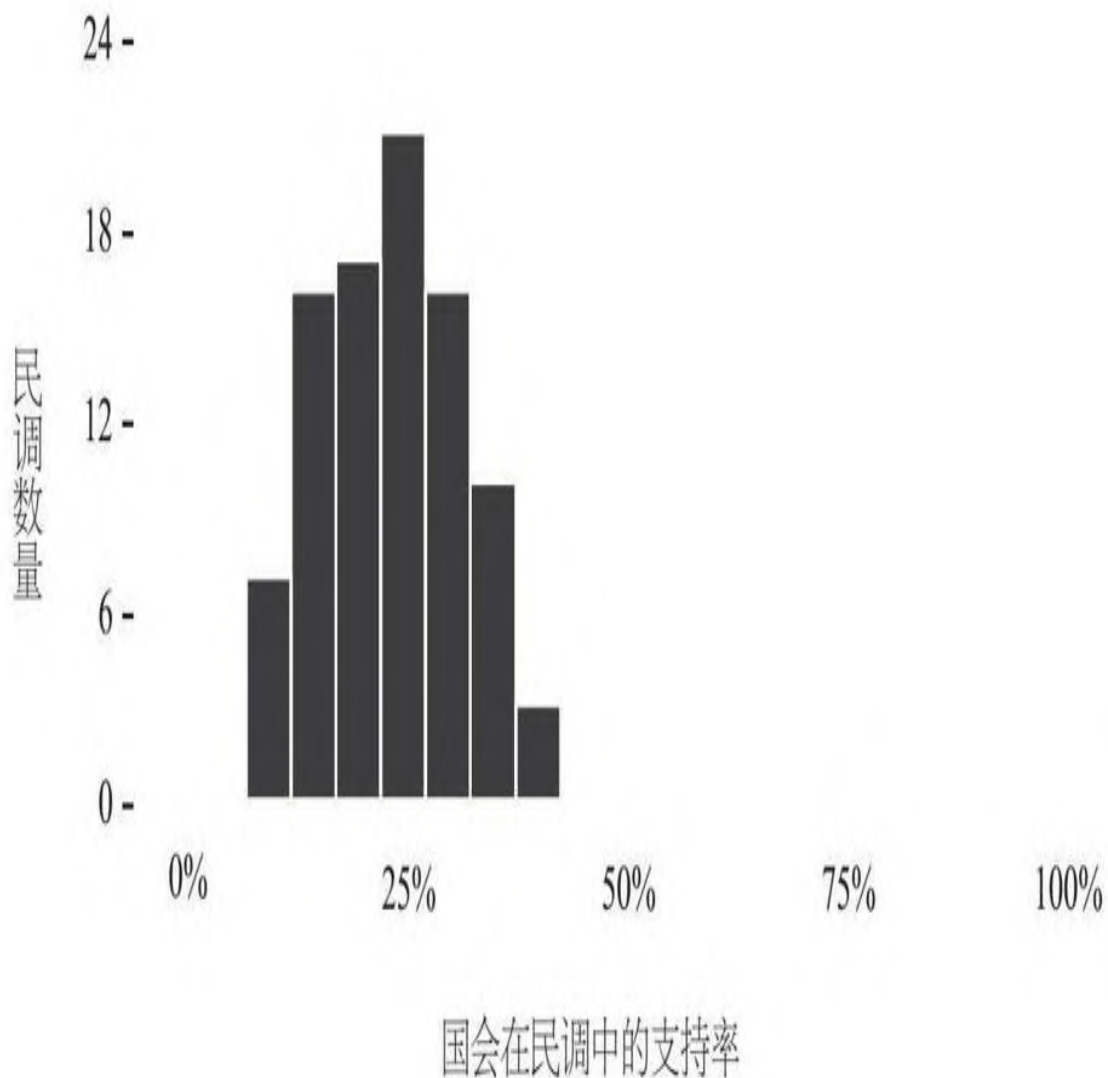


圖0—3 2008年11月以來，國會在主流媒體民調中的受歡迎程度

這些數據告訴我們，教練認為某個球員要麼非常優秀，有資格參加所有賽事的首發陣容，而有些球員則不優秀，沒有資格參加任何一場賽事的首發陣容。認為一個人要麼優秀，要麼不優秀，這是一個非常明顯的二分法。我們可能認為男性對女性外貌的評價數據也存在類似的情況，即如同頂級籃球天才和不優秀的籃球選手一樣，男性認為女性要麼美，要麼醜。但我們根據評價數據繪製的曲線圖卻揭示了一番不同的情景。通過數據理解事實往往不符合主觀認識，有時候面對無窮多的數據，將所有數據放在一起分析，更有可能揭示出偏離事實的大趨勢。但事實上，我們根據網站用戶的評價數據繪製的圖形非常具有對稱性，

類似於貝塔分佈曲線。人們通常用貝塔分佈曲線來模擬基礎性的、客觀公正的決策。下面就是我根據OkCupid網站的男性用戶對女性用戶外貌的評價情況繪製出來的曲線圖（見圖0—5）。

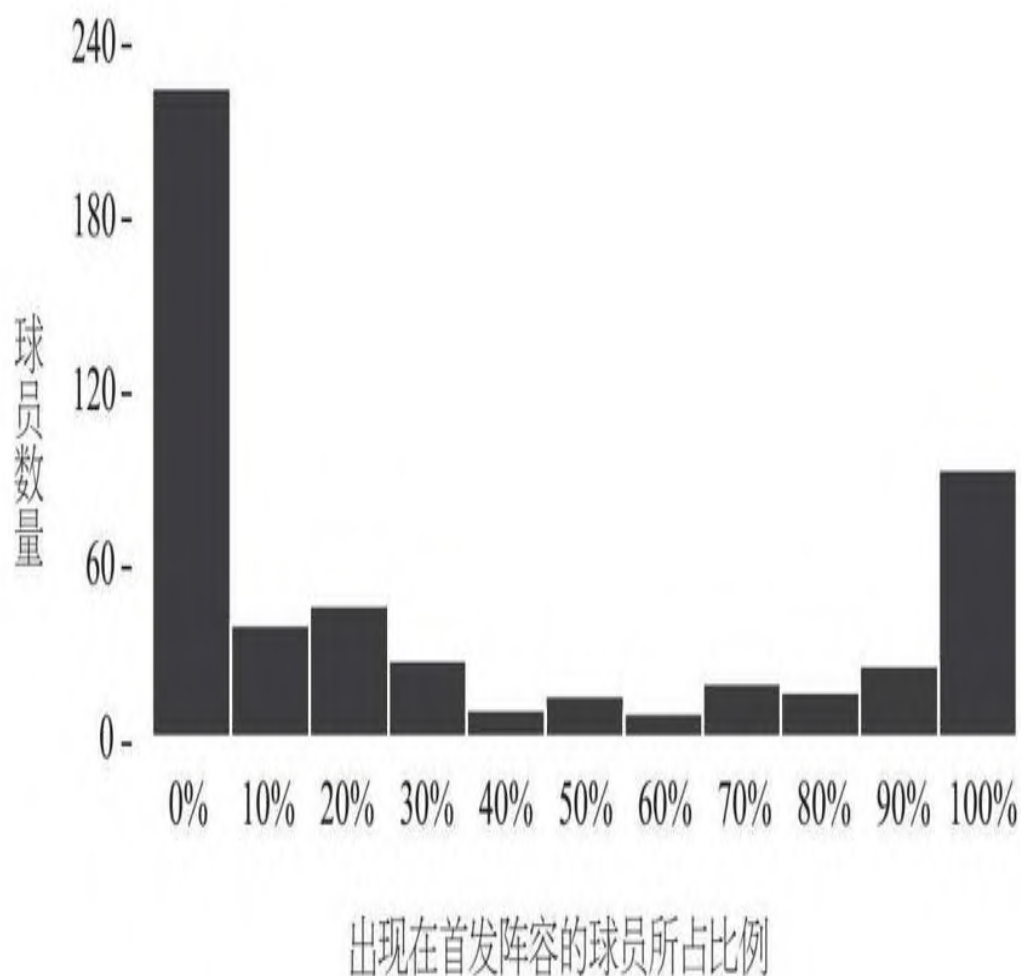


圖0—4 2012—2013賽季出現在首發陣容的NBA球員所佔比例

這幅曲線圖非常接近現實（偏差幅度在6%左右）。^[6]其實，我們即便憑空猜測，也可以猜到這樣的結果。很多教材就舉了類似的例子，我本人對這類例子向來都不放在心上。這幅曲線圖很好預測，集中度也比較明顯，大部分女性都集中在中間部分，因此，看上去似乎非常單調，令人感覺乏味。不過，這又有什麼關係呢？在這種情況下，令人乏味的圖形反而具有特殊的價值。這意味著做出這些評價的男性也如同這個圖形一樣具有可預測性和集中性，最重要的是，男性的評價具有公正性，不會出現大幅偏差。當男性看到名模、豔星、封面女郎、百威淡啤

廣告中的女郎以及用圖片處理軟件做出來的女性美圖時，男性的審美視角就會發生一定變化，永遠保持不變的審美視角可以說是一個奇蹟。男性對女性的外貌具有一些不切實際的期待，這幾乎是人所共知的常識。但與女性比較起來，男性在評價外貌方面比較慷慨，而女性則比較苛刻。我根據網站的女性用戶對男性用戶外貌的評價，繪製出了下面這幅圖（見圖0—6）。

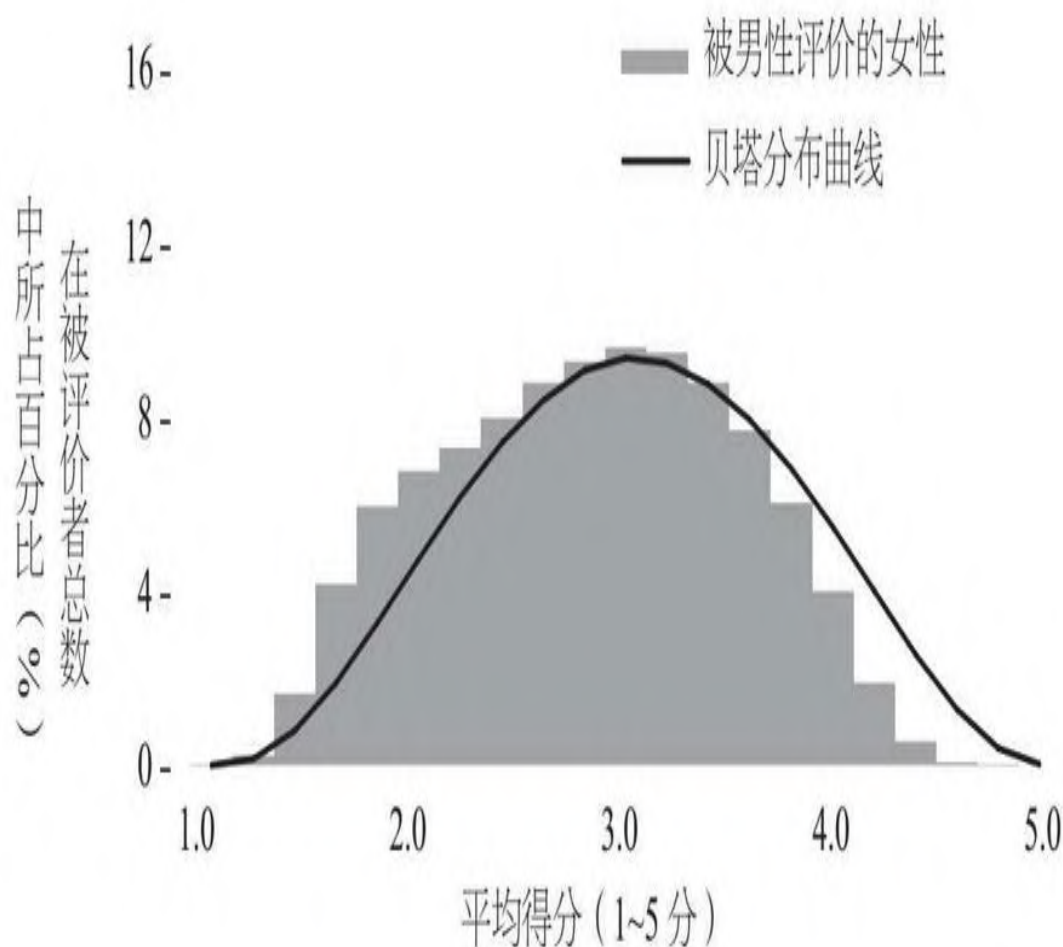


圖0—5 對女性魅力的看法

黑色部分的圖形幾乎集中在前1/4的部分。根據女性的評分，每6個男性中，只有1個人的得分超過了平均水平。人們往往不習慣於用這種量化方式來分析一個人的性魅力，所以，請讓我換個方式來分析。如果讓女性評價男性的智商，那麼女性認為58%的男性的腦子都壞掉了。[\[7\]](#)現在，OkCupid上的男性用戶其實並不像女性用戶評價的那麼醜。

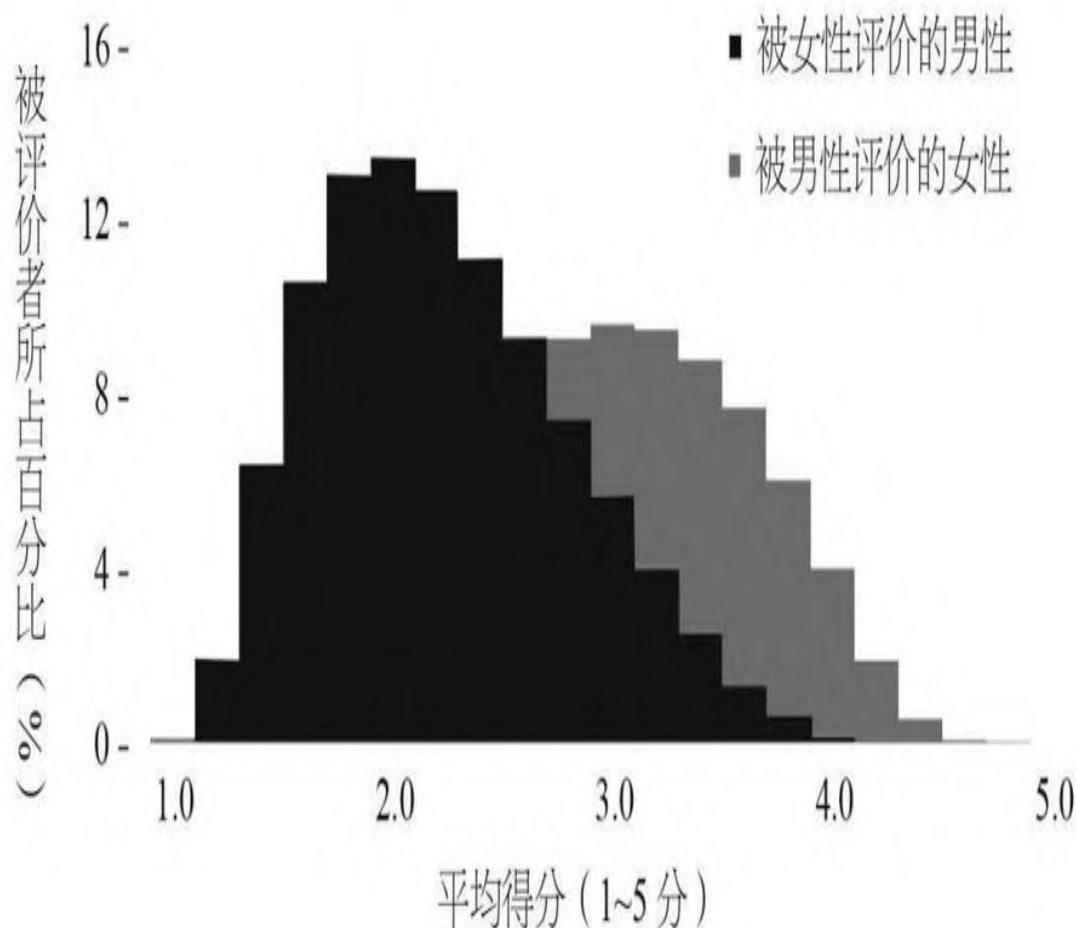


圖0—6 女性對男性魅力的評價與男性對女性魅力的評價

我專門做過這方面的對比實驗，然後得出了這個結論。從我們的用戶中隨機挑選出一組男士，然後從另外一個社交網站中隨機挑選出一組男士，將他們做對比。事實上，在我見過的每一個交友網站上，女性對男性魅力的評價情況都是相似的，都處於低位。Tinder、Match.com、DateHookup這三個交友網站的用戶在美國單身人士中所佔比重大約為50%。^[8]這就證明，男性和女性在評價對方性魅力時採取的細分標準是不同的。

《哈潑斯》（*Harper's*）雜誌曾經給出過一個說法，在性這個問題上，女性總是因為經歷過才感到後悔，而男性總是因為沒有經歷而感到後悔。這一點從上面男性和女性給予對方的評分情況就可以窺見一斑。我想補充一下，上面那些男性肯定會因為給予女性評分過高而感到後悔不已。^[9]

對於貝塔曲線，我們可以將其視為多次拋硬幣之後得出的結果。^[10]這種曲線反映了眾多相互獨立的二項事件發生的概率。在拋硬幣的過程中，正面朝上的概率與背面朝上的概率是相同的。但在我們的數據中，我們可以看到，女性對男性魅力的評價並非平均分佈的，而是集中在低分區，就相當於每拋四次硬幣才有一次正面朝上的情況。包括天氣在內的大量自然現象都可以用貝塔曲線來衡量。幸虧一些熱衷於研究天氣的人記錄下了歷史上的天氣模式，我才可以將人與人之間的評價同歷史上的天氣模式進行對比分析。男性對女性的評價標準就像用來預測紐約市雲量的函數一樣，屬於貝塔分佈曲線。而根據同樣的標準，女性的心理則像用來預測西雅圖雲量的函數一樣，喪失了平衡與對稱。^[11]

我們將根據前面這個思路來分析關於人際關係的數據。這類數據是本書將要分析的三類數據之一。我們將首先分析性魅力，看看性魅力究竟發生了哪些變化，以及究竟由哪些因素構成。我們將從技術角度出發，看一看為什麼女性魅力在21歲時達到了頂點以及明顯的文身的重要性，但這些與人體有關的數據很快就會分析完。之後，我們將以Twitter為數據庫，分析現代社會交往方式的變化，並看一看Facebook上的朋友圈對婚姻穩定性的看法。人們將頭像上傳到互聯網可以說是一件利弊並存的事情：它們幾乎把每一個網站（比如Facebook、招聘網站以及交友網站）都變成了選美秀場。OkCupid曾經做過一個實驗，將所有的用戶頭像刪除了一天，我們想看看這一天內會發生什麼。雖然我們往往認為愛情不是盲目的，但我們發現的一些證據卻表明，愛情竟然真的具有盲目性。

在第二部分，我們將分析關於人類隔閡的數據，首先是深入分析關於種族的數據。種族是人類社會存在的主要隔閡之一。現在，在互聯網的幫助下，我們有史以來第一次有機會將調查範圍擴大到絕大部分人類，並採集到真實的數據。作為交友網站的運營者，我們非常有幸擁有了這些數據。這些數據揭示了人們在種族問題上的態度，我們將看到，種族偏見不僅仍然十分強烈，而且是長期存在、根深蒂固的，但在公開場合，絕大多數人可能永遠都不會承認。幾乎在任何一個網站上，關於種族偏見的字句俯拾即是。一個人會把種族偏見視為祕密，隱藏在內心，但當自己覺得沒有別人觀察自己的時候，卻藉助鍵盤表達了出來。對於美國而言，種族主義可能是民眾最不喜歡的一個詞。我們將看一看谷歌搜索引擎為我們揭示了哪些內容。我們還將分析一下人的外貌產生的社會影響力，關於這一點，我們收集的數據是非常真實而具有說服力

的，比之前通過其他方式收集的數據強幾千倍。醜陋會給人造成驚人的社會成本，對於這些成本，我們現在終於也能進行量化分析了。我們還將根據Twitter的數據分析人類的易怒傾向。社交網站能夠讓我們在每一分、每一秒保持密切聯繫，同樣，也能讓我們在下一分、下一秒就斷絕聯繫。互聯網容易導致「集體憤怒」，從而引發新的暴力傾向，催生一批網絡暴民。可以說，暴民群體是世界上最古老的一個群體。

在第三部分，我們將會看一看關於人際交往的數據，其中既有正面的，也有負面的。我們將把目光集中到個體層面，探討社會成員如何表達自己的種族、性別和政治認同感，關注人們在自我表達過程中使用的文字、圖像和文化標記。白人女性在描述自己時，經常用下面這5個詞語：碧眼、紅髮、愛車一族、鄉村姑娘、戶外一族。

我們是分析美國著名鄉村音樂女歌手凱莉·安德伍德（Carrie Underwood）的歌詞呢，還是分析在網站上創建的數據呢？這是一個很好的問題。我們不僅要分析人們在公開場合所說的話，還要看一看人們在私下場合說了什麼和做了什麼，以便分析人們言論與行為之間的差異。我們還將看一看人們對爭議話題的看法，比如對雙性戀者的看法。這些話題可能會影響我們的身份認同。我們的數據來源是比較廣泛的，包括Twitter、Facebook和Reddit，甚至包括Craigslist網站，以便更好地瞭解我們在自己家裡的言論與行為，既有身體方面的，也有其他方面的。在進行廣泛深入的分析之後，我們就會自然而然地遇到這樣一個問題：在這樣一個一切都有可能遭到洩露的世界裡，人們如何保護自己的隱私呢？

在這本書中，我們將會發現互聯網是一個充滿活力、戾氣、關愛、寬容、詭詐、慾望和憤怒的地方。雖然網絡世界是由人類構建的，但將所有這些信息綜合起來去分析時，我敏銳地察覺到，這些數據並沒有捕捉到每個人的生活，因為如果你沒有電腦或智能手機，你便無法進入網絡世界。我意識到了這個問題的存在，但我會努力蒐集現有的數據，並等待這些人學會使用互聯網和進入網絡世界。

與此同時，我可以說Twitter、Facebook等網站的數據，甚至OkCupid網站的數據，絕對是真實的。如果你沒有親自使用這些服務，那麼你可能無法理解這一點。大約87%的美國人都使用網絡。^[12]這個數字幾乎適用於所有群體，從城市到農村，從富人到窮人，從非洲裔到亞裔，從白種人到拉美裔，基本上都被互聯網聯繫在了一起。^[13]只是在老年群體和受教育水平較低的群體中，互聯網的使用者所佔的比例較低

（約為60%）。在本書中，我將分析對象的年齡上限劃定在了50歲，這個年齡距離老年的認定標準還很遠，但我沒有考慮互聯網使用者的教育水平。超過1/3的美國人每天都會登錄Facebook，該網站在全球擁有13億個賬號。^[14]大約1/4的世界人口不滿14歲，而且大約1/4的成年人都擁有Facebook賬號。在過去三年內，本書中提到的交友網站的新註冊美國用戶達到了5 500萬人，美國每兩個單身人士中，就有一個擁有交友網站的賬號。從人口統計學上來看，Twitter是一個特別有趣的案例。它在技術領域取得了令人矚目的成就，這個公司幾乎憑一己之力讓舊金山的很多人躋身中產行列，但它提供的服務基本上是平民路線的，其交流平臺具有開放性，用戶背景也極為廣泛。例如，用戶不存在顯著的性別差異。^[15]只有高中教育水平的用戶與具有大學教育水平的用戶發佈的信息的數量是一樣的。拉美裔與白人的使用量也是相同的，而黑人的使用量則是他們的兩倍。當然，我們還分析了谷歌。如果說有87%的美國人使用互聯網，那麼肯定有87%的人使用谷歌。

這些龐大的數據並不能證明我的數據反映了全景。雖然仍有一些數據沒有囊括進來，但這些數據至少非常接近於全景。我的數據雖然並非完美無缺，但至少比以往任何時候的調查數據都可靠。比如，本書引用的數據量比蓋洛普或皮尤研究中心的調查數據多出了數千倍。這是不言自明的。不太明顯的一點是，事實上，我的數據比大多數學術性質的行為研究數據更具有包容性，因為與常規調查相比，網站用戶的背景與經歷更多元。

當前，行為科學領域的一些調查存在一個眾所周知的問題，即幾乎所有的基本觀念都是通過調查少數大學生得出的。這個問題很少被公開討論。當我還是一名大學生時，就有馬薩諸塞州總醫院的調查者支付給我25美元，讓我吸入一種具有輕微放射性的氣體，吸了一個小時，然後讓我躺進一個類似於電腦斷層掃描儀的機器裡面，按照他們的指示執行一些需要運用腦力來完成的任務，他們通過掃描獲得我的大腦圖像。他們告訴我，這就像在飛機上待了一年時間一樣，不會對我造成傷害，沒什麼大不了的。但當我躺在那個掃描儀器裡的時候卻感到有些難受。他們讓我閱讀文字，並用腳去點擊按鈕。但有一件事情當時他們沒有明確說明，我當時也沒有意識到，即他們是讓我代表白人男性去參與研究的。我的一個朋友也參與了這個研究。他像我一樣也是一個白人大學生。我敢打賭，大部分受試者和我們一樣都是白人。由於受試者的選擇範圍過於狹窄，我們遠遠無法代表所有男性，這種實驗得出的數據當然也不具有典型性和普適性。

我知道為什麼會發生這種情況。在現實中開展的實驗中，要得到一組真正具有代表性的數據，往往難度比較大。如果你是一位教授或博士後，希望推動一項研究，那麼你很可能傾向於選擇所謂的「便利樣本」，這就意味著你往往會選擇自己那所大學的學生。但這樣一來就導致「受試者範圍過於狹窄」的嚴重問題，導致你的數據不具有普適性，尤其是當你研究與信念、行為有關的內容時，這種問題更嚴重。這類調查研究甚至有一個專門的名字，即「怪異型」（WEIRD）研究。^[16]目前，大多數公開發表的社會調查類論文都屬於這一類。所謂「怪異型」研究，指的是絕大多數調查樣本都具有以下幾項特徵：白人（white）、受過良好教育（educated）、來自工業國（industrialized）、來自富裕國家（rich）以及來自民主國家（democratic），但符合這些特徵的人口僅佔全世界人口的12%。這5個英文單詞的首字母連起來就是WEIRD，意為「怪異的」。以在校大學生作為調查對象時，調查結果的怪異性尤為嚴重。^[17]

我的數據也存在部分上述問題，比如，我的數據曾經也過於注重來自工業國的人，過了很長一段時間之後才逐漸扭轉這個傾向。但由於技術往往被視為屬於「精英領域」，很多技術領域的人才都非常願意鼓勵社會調查更加註重技術人才，但我認為必須區分兩類技術人才：一類是企業家和風險投資家，他們經常出現在公共場合，大手一揮，滔滔不絕地發表自己的見解，他們的聲音充斥著我們的耳朵，他們的確非常符合「怪異型」人才的標準；另一類是技術類服務的使用者，他們中的絕大多數人是普通人，能夠代表普通大眾，他們的大眾性在一定程度上受到了所用網絡服務的影響，因為Twitter、Facebook、谷歌等類似網站提供的服務都是面向大眾的服務。

我可以肯定地說，從某種意義上來講，我收集的數據都是經過事實檢驗的，因為現在互聯網是人們日常生活的一個重要組成部分，這就在很大程度上確保了數據的真實性。以OkCupid網站的數據為例，為了交友，你在該網站上留下了你的城市、性別和年齡等個人信息，並在網站上留下對意中人的要求。只有當你留下了真實的信息時，網站才有可能幫你聯繫上合適的人喝咖啡或喝啤酒。你上傳的頭像應該是你本人的，是真實版本的自己。如果你上傳了一個更漂亮的人的照片或把照片美化得讓自己看起來比較年輕，那麼你的確有可能得到的更多約會邀請，但想象一下對方看到你之後的感受：他們肯定希望真實的你跟網上的頭像一模一樣。如果真實的你與頭像中的你存在鮮明的令人驚訝的反差，對方發現你沒有想象中的那麼好，那麼你的約會基本上就是「見光死」。

這個例子反映了一個大趨勢：線上世界與線下世界的融合會給互聯網帶來一種內在的社會壓力，抑制了很多互聯網用戶偽造信息的衝動。

交友網站、社交網站和新聞聚合器等網絡服務的用戶與現實生活中的絕大多數人一樣，都在人生之路上苦苦探索，只不過他們現在的探索工具增加了手機和電腦。他們幾乎在無意之中創造了一種獨特的檔案。這種檔案就是保存在網站服務器裡的數據庫。現在，世界各地都有這樣的數據庫，有的保存了人們在很多年間的願望和觀點，保存了不計其數的信息，以致看起來有些混亂。這些數據雖然存在了很多年，但沒有絲毫的損毀，極度精確。現在，對這些數據進行分析的時機已經成熟了。即便在10年前來看，這些數據的廣泛性和靈活性都是無法想象的。

我在過去幾年裡一直蒐集和解釋這些數據，不僅包括OkCupid網站，還包括幾乎每一個其他主要網站。然而，一個揮之不去的疑問一直讓我感到痛苦。我一直反對科技化和自動化，總感覺寫一本關於互聯網的書非常類似於畫一幅關於電影的畫。為什麼我這麼痛苦呢？這是在失意時刻引發的問題。

∞

《別回頭》（*Don't Look Back*）是一部非常好的紀錄片，忠實記錄了美國歌手鮑勃·迪倫（Bob Dylan）1965年在英國巡演的過程。我在大學時期看過很多次，我最好的朋友賈斯汀當時正在研究電影藝術。在這部紀錄片中有這樣一個場景：在一次酒會上，鮑勃和別人因為誰往街上扔玻璃瓶子的問題發生了爭執。顯然他們都喝醉了。衝突的高潮是下面這段對白，15年來這段對白一直印在我的腦海中。

迪倫：你看上去就像個惡婦，說話也像個惡婦。

對方：哎，去你的，你是個名人，你知道嗎？

迪倫：我知道自己是個名人，怎樣？

對方：我知道你有這個覺悟。

迪倫：我的名氣比你的大，夥計。

對方：我在你面前什麼都不是。

迪倫：是的。

後來，有人打破了他們的對話，然後他們又開始談論詩歌。這就是那個夜晚的一幕，但這一幕卻折射出這樣一個事實：人類社會迄今的歷史中似乎只記錄了名人的聲音，無論是搖滾明星還是其他名人。人類歷

史充斥著征服者、大亨、殉道者、救世主，甚至無賴。人類最早生活在幾條大河的沖積平原地區，而如今人類的足跡遍佈世界各地。從公元前3100年左右的埃及法老納爾邁（Narmer）到今天的史蒂夫·喬布斯和納爾遜·曼德拉，人類歷史的架構一直是以英雄主義為基礎的，敘事的內容往往是英雄如何號令世界。^[18]納爾邁是古代人類社會最早的國王之一。雖然人類的文字體系已經改變了，但這些國王的名字卻流傳了下來。即便到了20世紀60年代，這種情況仍然沒有改變，社會歷史仍然過於關注名人，而普通人的生活如果沒有與名人生活交叉的地方，就沒有什麼值得記錄的。比如，回眸那個時代，人們只是看到一些名人，包括保羅·麥卡特尼、約翰·列儂、鮑勃·迪倫、吉米·亨德里克斯等。

然而現在，這種不對稱的現象正趨於終結，因為除了名人的聲音之外，普通人的聲音終於也能被記錄下來了。隨著互聯網的發展，新聞、攝影、慈善、喜劇和很多其他領域的大眾化趨勢日益明顯，我希望這一趨勢最終也能影響到我們人類的敘事方式，在歷史敘事中更多地融入普通人的聲音。現在，這一趨勢剛剛開始顯現出來。我寫這本書的目的之一就是為讀者呈現我和其他人在這個方面的研究成果，但普通人的聲音仍然比較微弱，也沒有經過提煉，就像火車呼嘯而過之後留下的鐵軌顫動的聲音。數據科學還遠遠沒有發展到完美的地步，數據選取過程中仍然存在一些偏見，還存在許多缺陷，我們要理解、承認以及克服這些缺陷。但數據與現實之間的差距日漸縮小，最終必然會完全一致。

關於大數據，我知道肯定會有很多人發表一些宏大的言論，但我並沒有說大數據會改變歷史進程。大數據肯定不會像內燃機或鋼鐵那樣改變歷史進程，不過我相信大數據會改變人類對歷史的認知，讓人們重新考慮究竟什麼是歷史。有了大數據之後，歷史會變得更加厚重、更加廣博。在歷史上，人類用來記載歷史的載體無非是泥版、莎草紙、新聞紙、賽璐璐片、相紙或其他材料製成的紙。與這些載體相比，記載數據只需要利用硬盤就行了。硬盤空間不僅非常廉價，而且幾乎用之不盡。硬盤有足夠的空間，不僅能容納英雄，還能容納普通人。事實上，我本人也算不上一個英雄，只是一個小人物，喜歡與朋友和家人聚在一起，過著普普通通的生活。因此，對我而言，大數據時代的到來是很有意義的。

現在，雖然我希望我、你和其他許多普通人也能與總統們一同被歷史記錄下來，當後人研究這個時代時能發現我們普通人的身影，但肯定會漏掉一些人，即便最好的數據也改變不了這一點。但我們終將被記錄

下來。10年、20年或100年之後，如果有人研究這個時代，想要理解這個時代的變化，如同性戀婚姻合法如何推動和折射了社會對同性戀的接納程度，以及亞洲城中村的衰落與重生，那麼他們研究的依據將會是來自Facebook、Twitter、Reddit等網站的數據。不然的話，當前我們這些在網站上寫下各種數據的人就算失敗了。

我試圖通過本書的標題傳遞出大數據時代的特徵。在《舊約》中，希臘語Kataklysmos的意思是「洪水」「災難」，英語中的cataclysm（洪水、災難）一詞就是由這個詞演變而來的。我在cataclysm的基礎上發明了dataclysm一詞，這個詞有雙重含義。第一重含義是，在大數據時代，數據的規模之大是前所未有的，如同滔滔洪水一般向我們襲來。我們今天能夠收集的數據廣泛而深入，幾乎是無限的。如果說傳統上那些調查研究收集的數據是毛毛雨，那麼我們今天通過網絡收集的數據無異於連續下了40個晝夜的傾盆大雨。第二重含義是，我希望洪水般的大數據能夠沖走人類往日的狹隘思維和今天的有限視野，從而推動這個世界的變革。

這本書是由一系列的小插曲組成的，每一個小插曲就是一扇小窗，通過這一扇扇窗，我們能夠觀察自己的生活，看看哪些因素會讓我們會聚在一起，哪些因素會讓我們產生隔閡，以及哪些因素造就了當前的自己。隨著數據源源不斷地湧來，我們觀察自己人生的窗口會越來越大，而且最早的觀察是最令人激動的。接下來，我將帶你走近這些窗口。

[1] 為了得出這個數據，我統計了2014年4月之前的12個月內的訪客數量，具體數字為10 922 722。

[2] 對於在線約會網站的經營者而言，有多少用戶會真正約會，以及約會後會發生什麼事情，是很難準確得知的。我在這一段所講的內容是自己對用戶約會情況所做的儘可能合理的推測。我採用了兩種不同的推理方式。第一，我假設每個積極的用戶應該每兩個月約會一次，我覺得這種假設算是很保守了。由於我們每個月有400萬個活躍用戶，這就代表著每天大概有6.5萬人約會，相當於3萬對男女。第二，每天，大約300對情侶登錄「停用賬號」界面，表示自己已經用OkCupid網站找到了穩定的伴侶，再也不需要登錄這個網站了。這些情侶不僅確實在認真地約會，認真到了停用OkCupid賬號的程度，而且還不厭其煩地填寫一系列表格，將自己的最新感情狀態告訴我們。我估計，由OkCupid促成的長期情侶中，只有1/10的情侶會註銷賬號，並告訴我們最新的感情狀態。此外，我還估計，在首次約會後，變成這種長期伴侶的概率也是1/10。也就是說，一個人同10個人約會之後，才可能找到一個長期伴侶。換句話講，在每天通過我們網站進行約會的3萬對男女中，大概只能促成3000對長期伴侶，而每3000對長期伴侶中，最終結婚的比例估計只有1/10。我們不妨換個角度看待這個現象，即你在結婚之前，要認真地談幾次戀愛呢？我估計平均大約是10次。

[3] 通過Foursquare這款軟件公開發布的應用程序界面（API），隨機選取紐約市305家比薩店計算出平均得分。

[4] 這個數據來自2009年1月26日到2013年9月14日在realclearpolitics.com網站的529份對於

「國會工作滿意度」的民意調查資料。網址鏈接如下：
realclearpolitics.com/epolls/other/congressional_job_approval-903.html#polls。

[5] 該圖數據來源於espn.com網站，顯示了2012—2013賽季參加首發陣容的NBA球員所佔比例。沒錯，我統計的對象就是NBA的一支老牌球隊——費城76人隊。

[6] 這個數字是這樣得來的：先在曲線上取21個離散數據點，然後計算這些數據點距離的幾何平均數，計算結果是0.056。計算公式為：

[7] 以女性吸引力曲線的中間位置為例，男性吸引力曲線的中間位置比女性相差了整整一個標準偏差。如果將這種差距轉換成「智商」這一指標，就意味著男性智商的中位數低於85，而85是「邊緣智力」的門檻。舉個例子，美國軍隊就不接受智商低於85者入伍。「腦子壞掉」這個說法固然存在誇張成分，但其目的在於希望讀者感受到男女智商中位數的差距之大。嚴格來講，我們可以說，大約58%的男性的智商可能不到85。

[8] 如果讓我準確地說明我掌握的用戶數據的涵蓋範圍究竟有多大，肯定是一個挑戰。我努力用一些寬泛的、易於理解的措辭去表達自己的猜想，因為我知道約會網站不同於Facebook和Twitter，這本書的很多讀者從來沒有用過約會網站。比如，如果你在20世紀90年代晚期或者之前就已經結婚了，或者找到了穩定的伴侶，那麼你就不會用任何約會網站。根據2011年的人口普查數據，在15~64歲的美國人裡，有1.03億單身人口。但這個數字指的是所有未婚人口，其中包括很多擁有長期伴侶卻未結婚的人，也包括同性戀者。2011—2013年，Tinder、OkCupid、DateHookup和Match.com這4個在線交友網站在整個美國的註冊用戶達到了5 700萬。其中，僅僅在2013年這一年之內，新註冊用戶就多達2300萬人。我之所以說Tinder、Match.com、DateHookup這三個交友網站的用戶在美國單身人士中所佔比重大約為50%，是先計算出5 700萬與1.03億之間的比例，然後扣除大約10%~15%的重複賬號。

[9] 這個說法出自拉斐爾·克羅爾—查迪（Rafil Kroll-Zaidi）於2014年2月在《哈潑斯》雜誌「發現」（Findings）欄目發表的文章。

[10] 我的數據分析師湯姆·奎塞爾（Tom Quisel）幫我用非常淺顯易懂的語言解釋了具有二項分佈特徵的貝塔曲線。他還指出貝塔曲線可以用於建立天氣預報模型，並將這個曲線與weatherbug.com上面各個城市的天氣變化曲線做了對比。

[11] 西雅圖有「雨城」之稱，其陰天時間每年平均為220多天，遠遠多於紐約的130多天。——譯者注

[12] 關於這個數字的來源，請參考皮尤研究中心的蘇珊娜·福克斯（Susannah Fox）與李·蘭尼埃（Lee Rainie）撰寫的互聯網研究項目的成果摘要，時間為2014年2月27日，網址為：pewinternet.org/2014/02/27/summary-of-findings-3/。

[13] 舉個例子，白人、非裔美國人和拉丁裔美國人的互聯網使用率分別是85%、81%和83%。據此推測，亞裔美國人的網絡使用率應該相差無幾。關於年齡問題，除了65歲以上的人之外，各個年齡段人群的網絡使用率都在80%以上。詳情請參考皮尤研究中心的蘇珊娜·福克斯與李·蘭尼埃撰寫的互聯網研究項目的成果摘要，時間為2014年2月27日，網址為：org/files/2014/02/12-internet-users-in-2014.jpg。

[14] 2013年8月，Facebook在其報告中宣稱擁有1.28億美國用戶。2013年9月，Facebook在全球至少擁有12.6億用戶。全球人口總數以及美國人口總數則來自維基百科，請參考：expandedramblings.com/index.php/by-the-numbers-17amazing-facebook-stats/。

[15] 對於任何研究過社交媒體的人而言（不能僅僅關注谷歌眼鏡引發的爭議），這一點差不多是基本常識。請參考皮尤研究中心發佈的題為《主要社交網絡平臺的人口學特徵》的報告。該報告表明，「高中及以下學歷用戶」與「大學及以上學歷用戶」在Twitter用戶總量中所佔的比例基本上不存在統計學上的重大差異（二者分別是17%和18%）。皮尤研究中心選取的美國18歲及以上人群的方式是隨機的。從種族角度來看，黑人用戶所佔比例為29%，白人和拉

丁裔美國人用戶的比例均為16%。該報告的作者為梅芙·達根（Maeve Duggan）和阿龍·史密斯（Aaron Smith），完整報告的獲取網址如下：pewinternet.org/2013/12/30/demographics-of-key-social-networking-platforms/。

[16] 這一事實以及我對這一事實的態度源自貝薩尼·布魯克謝（Bethany Brookshire）於2013年5月8日發表的《怪異心理學》（Psychology Is Weird）一文，網址鏈接為：http://www.slate.com/articles/health_and_science/science/2013/05/weird_psychology_social_science_research.html。另外請參考《經濟學人》雜誌於2012年5月24日刊登的《民眾的咆哮》（Roar of the Crowd）一文，網址鏈接為：economist.com/node/21555876。

[17] Slate（《石板》）雜誌曾經指出：「參與怪異型研究的受試者來自的國家只能代表全球大約12%的人口，這些人口的道德決策、推理方式、公平理念，甚至視覺感知都不同於其他人口，因為很多行為與感知都是以我們賴以成長的環境為基礎的。」

[18] 你可以想象得到，這一點肯定是存在爭議的，但納爾邁無疑是一個可能性比較大的人選。在之前的書稿中，我曾經認為吉爾伽美什（Gilgamesh）是人類今天準確得知姓名的最早的人，因為J.M.羅伯茨（J.M.Roberts）在其所著的《世界史》（History of the World，牛津大學出版社，1993年）中提出了這種觀點，我受到了他的影響。但由於考慮到納爾邁所處的那個時代早了7個世紀，而且我覺得他可能也是真實存在的人物，所以我最終決定選擇納爾邁作為我們今天準確得知姓名的最早的人類。在「雅虎知識堂」（Yahoo! Answers），甚至有人戲謔地說，人類今天準確得知姓名的最早的人是美國著名搖滾明星埃爾維斯·普雷斯利（Elvis Presley）。

第一部分 我們因何而聚

第一章 伍德森法則

在世界上一些險峻的地方，比如說安第斯山脈，人們出行時會選擇軌道纜車——高山上的滑輪連接著一對纜車。一輛纜車下行時產生的動能將另一輛纜車拉上山，兩車通過平衡力運行。我逐漸瞭解到為人父母就是這種感覺。如果歲月讓我日漸衰老，那麼歲月也讓我的女兒茁壯成長。若是這樣，就讓一切順其自然。當然，聽任時光流逝，我滿心欣喜，特別是每個流逝的瞬間，也見證著我和女兒多一分的相互陪伴，但這並不意味著我不懷念過去，那時我未生華髮，皮膚也沒長出難看的斑點。我的女兒現在還小，我敢說，手背上的皺紋最能體現時光的流逝，我握住女兒胖乎乎的手指，教她數數：1，2，3……

然而有了孩子，開始長皺紋並不是什麼新鮮事。你從使用玉蘭油市場部本週主推的護膚品開始——我指的是用淺褐色面霜來「修正膚色」，這些面霜不是用阿爾薩斯地區山地的泥巴做的，就是用什麼其他瞎扯的東西做的——持續使用，直到青春容顏再現。自從有了人類以及「抗拒」和「醜陋」這些概念，人們就一直抗拒變老，並痴迷於抵制因老而醜。「死亡」和「納稅」是人生不可避免的兩件事，對吧？若是指望著下一次政府停擺，納稅看似越來越不可靠了。那好吧，就這樣吧。

在我還是個青少年的時候——我突然意識到，比起現在的我，當時的我更接近女兒的年齡——我很喜歡朋克搖滾樂隊。我現在喜歡的樂隊是比較傲慢卻不那麼專業的綠日樂隊。現在回過頭來聽那些樂隊的歌，整個感覺似乎是超自然的：似乎有什麼不可見的力量，把三四個成年人聚在一起唱歌，來抱怨自己的女朋友或是其他人在吃什麼之類的事。但在當時，我覺得這些樂隊無比炫酷，並且因為他們太過炫酷，所以連張海報都沒有，我只能在臥室的牆上貼上他們的專輯封面或是宣傳單。從那以後，父母早就已經搬家了——事實上，搬了兩次。我很確定我的舊臥室現在已經是別人的閣樓了，我也不知道我收集的那些東西去哪裡

了。我甚至想不起來它們的樣子，我只記得我有過這些東西，想到這裡，我微微露出笑容，又感到不安。

但今天，如果一個18歲的少年把一張圖片上傳到他的社交網站網頁上，這張圖片永遠都不會遺失，不但他本人在38歲時能回首這一天，拾起昔日時光的碎片，並自問「我當時在想什麼」，我們也可以這樣做，研究人員也可以。不但如此，研究人員讓所有人，而不僅是這位少年，實現這種對過去的回顧，而且研究人員可以把那個少年的18歲和18歲之前以及之後的時光連接起來，因為那面滿是各種圖畫和照片的牆，將一直追隨著他。從家裡的臥室到學校的寢室，從自己的第一間公寓到女朋友的房間，從度蜜月的房間到女兒的育兒室，他會繼續在牆上更新女兒吃米粉的照片。

新晉父母可能是對人生里程碑最為敏感的人。他們和別人聊天幾乎只聊這個，並且每隔幾個月去看醫生時，會了解兒童生長髮育指標。但是在寶寶中心網站（babycenter.com）和兒科醫生不再提醒父母很久之後，新的里程碑還是會一直出現，只不過我們不再記錄這些里程碑。然而，用電腦記錄是最合適不過的了；記錄進展是電腦唯一的工作。電腦不會遺失剪貼簿，不會出門旅行，不會醉酒，不會變老甚至不會眨眼，電腦就在那裡記錄。生命中的無數時期，一旦走過，就只留存在記憶中或只剩一些紀念品放在鞋盒中保存；而現在，生命的各個時期可以被長期保存，儘管這會嚇到Instagram裡有醉酒自拍照的人，但是如果處理得當，不言而喻，這是一個可以增進理解的機會。

如上述所言，記憶牆以及人生的長時間積累，正是社會學家口中的「縱向數據」——隨著時間流逝，收集有關同一個人的數據——以及我個人對未來研究的猜測。人類現在還沒有這種能力，因為互聯網作為一個普通的人類活動「記錄者」，還過於稚嫩。儘管難以相信，但Facebook和Touchstone等網站提供的服務也只能記錄6年時光。深度的信息資料是我們努力的方向。10年或者20年之後，我們可以這麼問自己，如果一個人自嬰兒時期起，生命的每個片段都被上傳，所有人都可以看見，這對那個人來說是多大的困擾？與此同時，我們也聽過很多關於朋友日漸疏遠、新理念滲入主流想法的故事。從中可以看出數據庫中數據的長遠潛力，我們也可以在其他地方看到這些數據的潛力，比如Facebook的TimeLine（時間線）功能：隨著時間的流逝，數據創造了一種新的圓滿，甚至是一門新科學。

現在在某些情況下，我們甚至可以找到完美的方式，超前描繪事物

發展的可能性。我們找來一組人，找出他們生命中不同時段的某個點，比較這些點，由此大致瞭解他們的生命軌跡。這個方法不適用於音樂品位，因為音樂本身會隨時間發生變化，所以分析時無對照組可言。但對於一些固定的、具有普遍性的事物而言，可以通過一些特徵來預測。美貌、性魅力和年齡就屬於這類事物，它們之間的關係是固定的，具有普遍性，因此是可以預測的。我擁有的數據可以對這些事物的演變做出預測。現在，在互聯網的幫助下，我們已經有可能通過分析部分人的數據來預測大趨勢，這在人類統計史上具有里程碑式的意義。但到目前為止，這種分析過程仍然會暴露出一些缺陷，以致無法完美無缺地反映事實。為了說明大數據時代的分析過程，我們接下來分析一下男性與女性對彼此美貌、性魅力和年齡的看法。在這個話題上，自古至今已經有無數作家、畫家、哲學家和詩人進行了深刻的分析，可謂眾說紛紜、莫衷一是。與他們的分析方法相比，我的分析方法似乎缺少了幾分藝術性，但多了幾分精準。思想與行動之間往往存在一些距離，而正是這段距離能夠揭示出一些精彩的內容。我將為你展示一下我們是如何發現這些精彩內容的。

我首先分析女性對男性魅力的看法。下面的所有數據都是真實的，是從多個網站蒐集來的，但為了便於講述，我只引用了OkCupid網站的數據。下面的數據反映了在女性眼中，哪個年齡的男性最有魅力。如果我用不同的方式來展示這些數據，你立刻就能看出為什麼會出現這種情況。

20	23
21	23
22	24
23	25
24	25
25	26
26	27
27	28
28	29
29	29
30	30
31	31
32	31
33	32
34	32
35	34
36	35
37	36
38	37
39	38
40	38
41	38
42	39
43	39
44	39
45	40
46	38
47	39
48	40
49	45
50	46

女性年齡vs女性眼中男性最有魅力的年齡

我們從上往下看，20歲和21歲的女性比較喜歡23歲的男性，22歲的女性比較喜歡24歲的男性；依此順序向下看，50歲的女性最喜歡46歲的男性。這並不是我調查得出的數據，而是根據人們在交友網站上留下的相關信息整理出來的。^[1]即便只是看前幾個年齡的女性及其偏好，就能清楚地發現數據傳遞出的信息：女性偏好與自己年齡相仿的男性。女性年齡低於40歲的話，其偏好的男性年齡往往非常接近於女性年齡。圖1—1中，左側數字表示女性年齡，虛線附近的數字表示女性偏好的男性年齡，那麼就得出了這幅圖，大趨勢就會更加明顯。

虛線對角線是一條年齡等位線，表示男女年齡是相同的。這不是一個經典的數學圖形，而是我為了引導你看出大趨勢而繪製的。通常來講，每一種情形都會有內在的幾何原理，幾何學之所以被稱為最早的科學，是有一定原因的。^[2]如果能利用幾何學來說明問題，我們就儘量利用其來更加清楚地揭示大趨勢。這條線為我們揭示出了兩個過渡年齡，這兩個年齡都屬於大齡。第一個是30歲。過了30歲之後，女性偏好的男性年齡，就一直位於對角線的左側，再也沒有回到右側。這就意味著30歲之前的女性，偏好比自己年齡略大的男性；而過了30歲的女性，則偏好比自己年齡略小的男性。第二個過渡年齡是40歲。40歲以上的女性偏好的男性年齡與女性年齡之間的差距開始拉大，甚至一度相差將近9歲。也就是說，無論你之前的想法如何，這些數據表明，40歲以上男性的魅力會迅速下降，40歲以上的女性對男性偏好變化非常大，開始喜歡比自己年齡小的男性。如果我們一定要找出男性魅力的轉折點，那就是40歲。男性過了40歲，在女性眼中的魅力就會迅速下降。

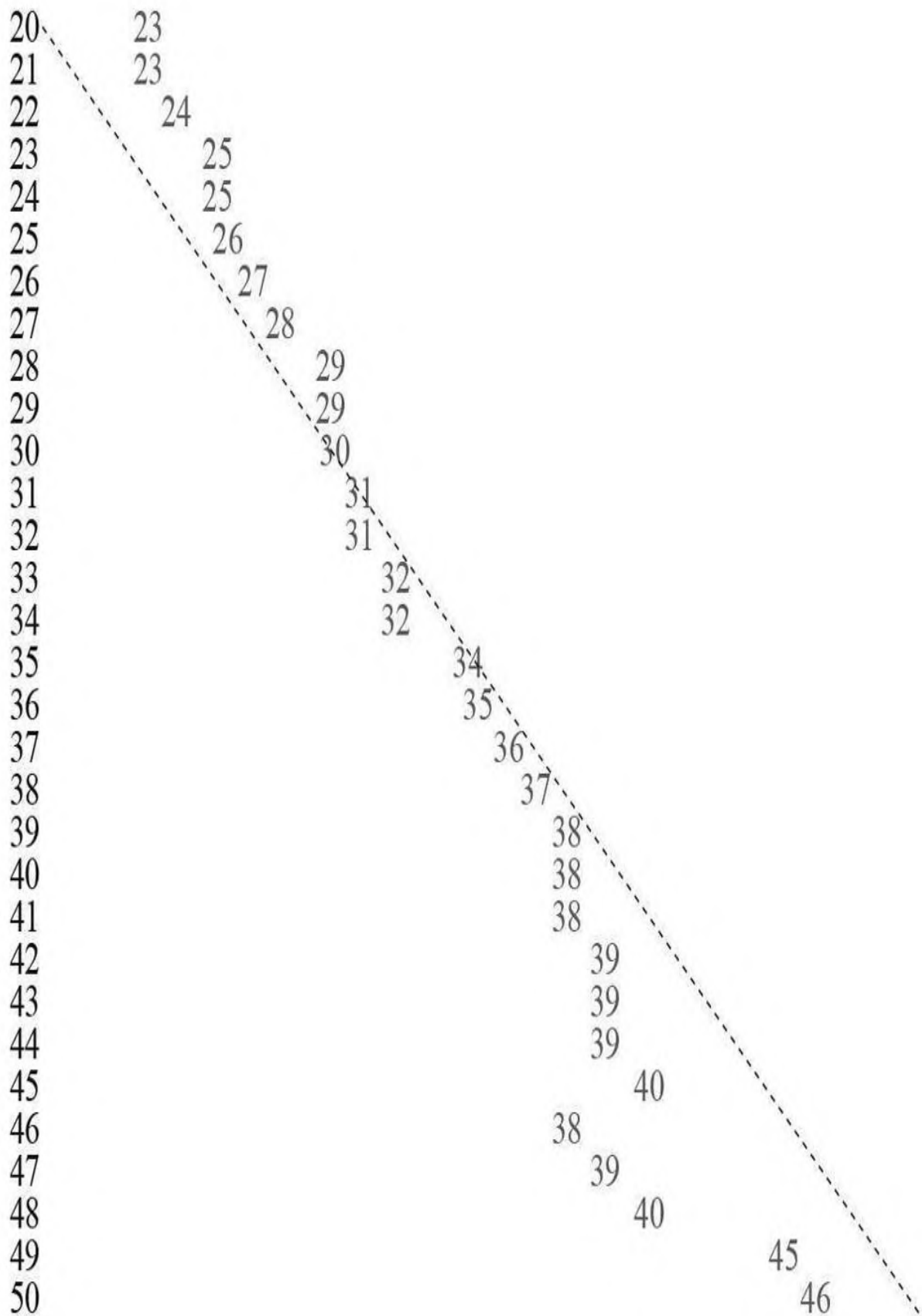


圖1—1 女性年齡vs女性眼中男性最有魅力的年齡

女性看男性魅力的方式與男性看女性魅力的視角存在天壤之別。隨著女性年齡漸增，其標準也會變化。黑色數字（即女性年齡）與灰色數字即女性認為男性最有魅力的年齡的變化趨勢大體上是一致的，這就意味著男性年齡越大，就越成熟；女性年齡越大，對男性的期待也越高。即便男性逐漸變老，皺紋和鼻毛會越來越明顯，而且會越來越熱衷於穿大口袋短褲，也總有一些女性認為這些無關緊要，或者男性會有其他魅力來抵消這些負面因素的影響。因此，年長的男性往往能得到類似年齡段的女性的接納。與此恰恰相反的是，男性對女性的審美視角則是比較固定的。用虛線對角線來表示的話，最有魅力的女性年齡的變化呈現出了自由落體的態勢（見圖1—2）。

圖1—2與圖1—1揭示出來的變化趨勢形成了鮮明對比。無論男性年齡如何增加，他們永遠認為20歲剛出頭的女性最有魅力。真實情況就是這樣，我繪製的這幅圖可能無法有力地揭示這個大趨勢。對於絕大多數年齡段的男性而言，女性最有魅力的年齡只有4個：20歲、21歲、22歲和23歲。只有45歲的男性認為女性最有魅力的年齡是24歲。從圖1—3中，你或許能更加清楚地看出男性審視女性魅力的視角是如何變化的。在圖1—3中，我根據男性對女性年齡的偏好程度進行了著色。為了便於讀者理解，我在橫軸上標上了女性的年齡，並將女性20歲到50歲平均劃分成了四個區間。

20 20

21 20

22 21

23 21

24 21

25 21

26 22

27 21

28 20

29 20

30 20

31 20

32 20

33 20

34 20

35 20

36 20

37 22

38 20

39 20

40 21

41 21

42 20

43 23

44 21

45 24

46 20

47 20

48 23

49 20

50 22

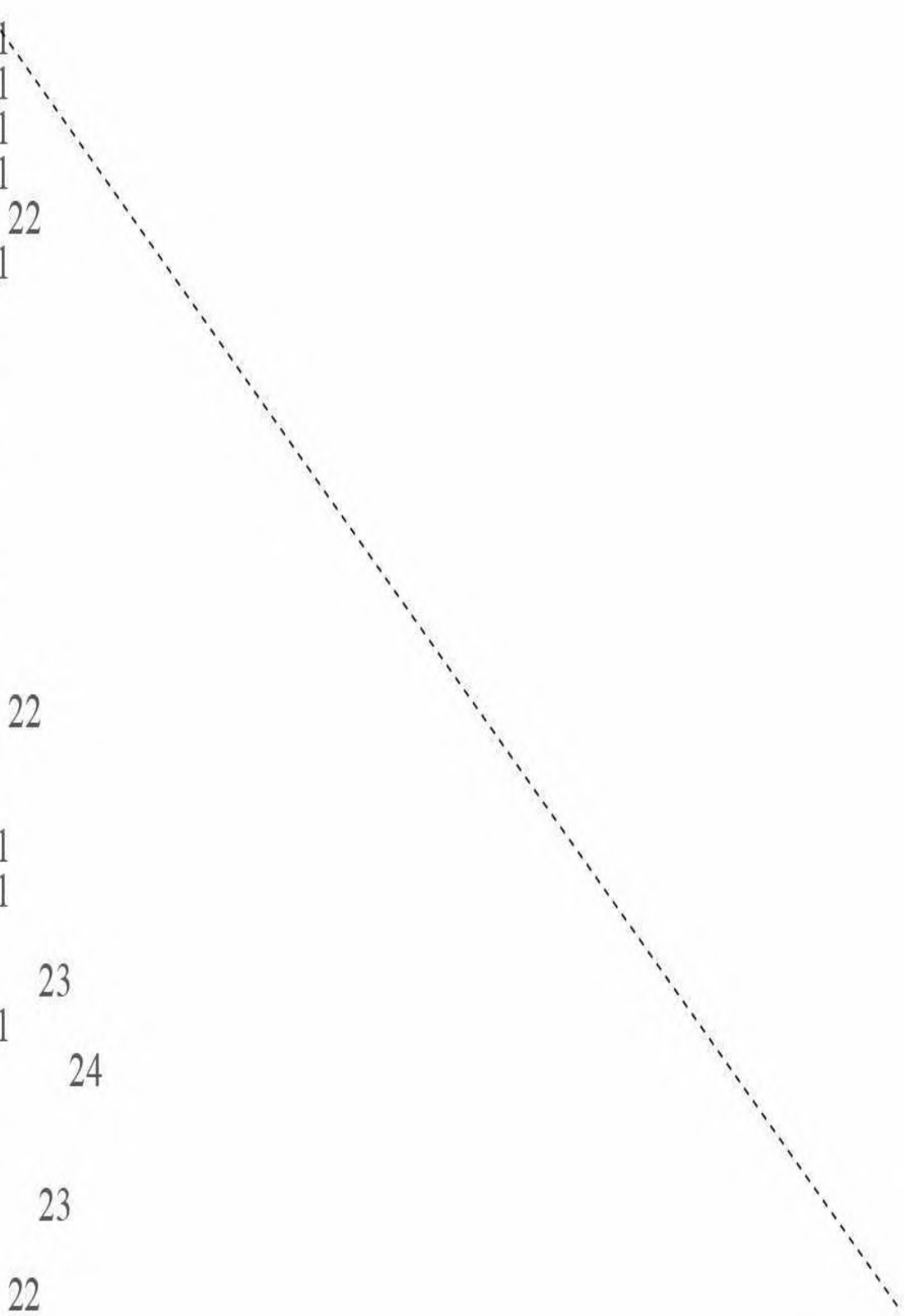


圖1—2 男性年齡vs男性眼中女性最有魅力的年齡

從圖1—3中可以看出，男性不僅僅喜歡20多歲的女性，而且男性在30歲之後，幾乎不會喜歡35歲以上的女性。在男性看來，女性越年輕越好，20歲左右的女性是最有魅力的。換句話講，女性達到法定的飲酒年齡時，魅力達到了最高值，之後便開始走下坡路了。

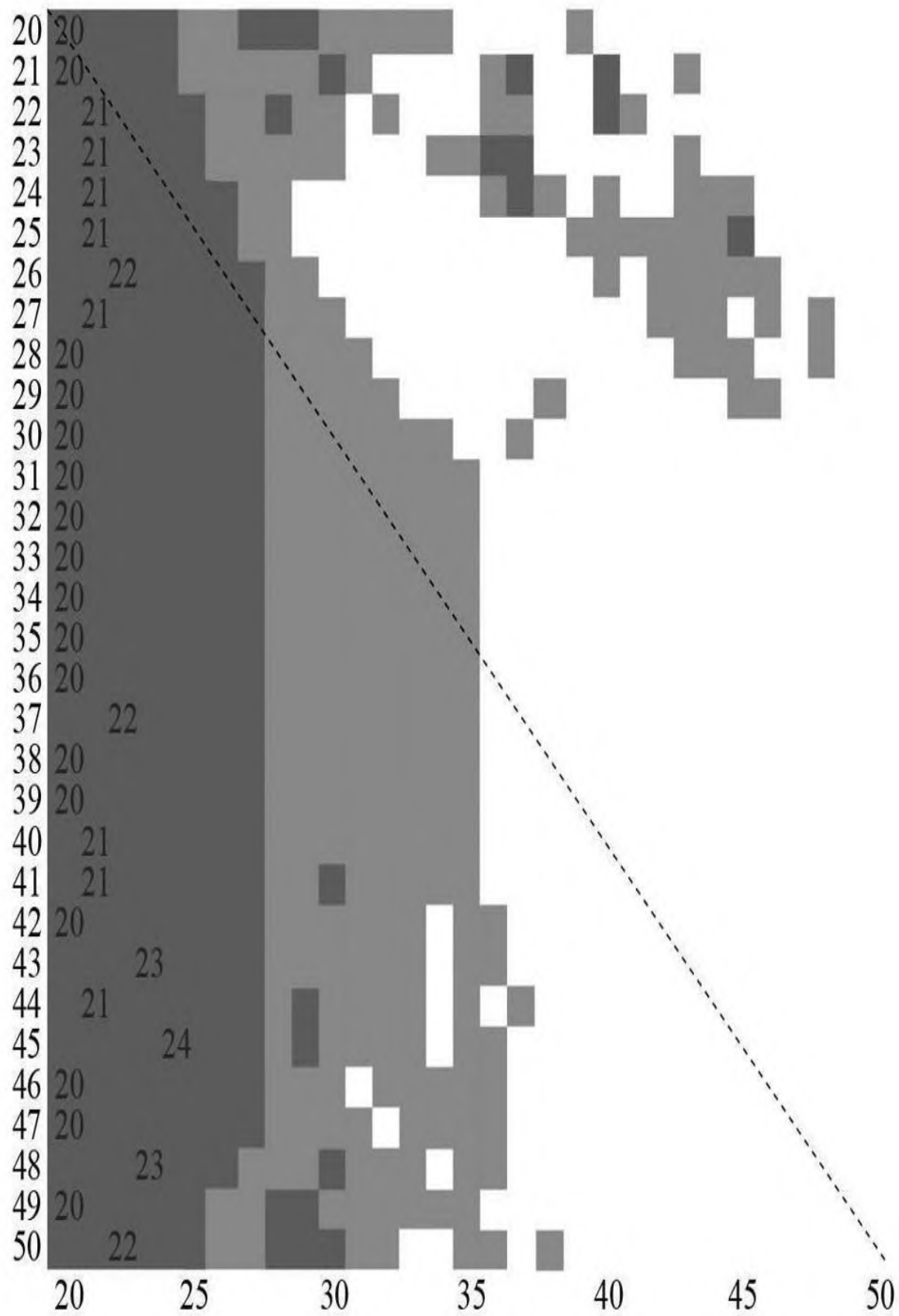


圖1—3 男性年齡vs男性眼中女性最有魅力的年齡

當然，對於男性過於關注年輕女性的另一種解讀是，男性對女性魅力的期望值永遠不會增長。一名50歲男性的審美視角與大學生的審美視角大抵相當，兩者都會把年齡作為考慮的一個變量，但20多歲的年輕男性更願意同年齡較大的女性約會。對角線右上角那部分陰影中的女性基本上都是「熟女」，年輕男性願意與這類女性約會，但一種常見的情況是當年輕男性與這類女性一起外出旅行、享受美好的一天之後，往往有一方發現自己被對方欺騙了，關係便戛然而止。

從數學角度來看，男性年齡與其女性目標是獨立變量，前者變化了，而後者卻不會隨之改變，各個年齡段的男性總是對20歲的女性最感興趣。我將這種法則稱為「伍德森法則」（Wooderson's law）。之所以取這個名稱，是為了紀念這個法則最有名的支持者馬修·麥康納

（Matthew McConaughey）。他在《年少輕狂》（*Dazed and Confused*）這部電影中飾演伍德森，伍德森有一句名言就是：「我就是喜歡高中女生。我變老了，她們的年齡永遠不變。」（見圖1—4）

然而，與伍德森不同的是，在男性的公開言論中，他們聲稱喜歡的女性類型卻不同於我們剛才看到的數據，也就是他們私下的評價數據。上面的評價數據是他們在沒有外人知道的情況下對女性的評價，但當你坦率地問男性他們正在尋找哪個年齡段的女性時，你會得到不我認為OkCupid網站的用戶在輸入自己的偏好時不會故意誤導我們，因為他們幾乎沒有這麼做的動力，如果輸入了錯誤的偏好，網站就會給他們推薦錯誤的人選，而他們知道自己肯定不會喜歡這些人選的。所以，這就迫使他們輸入正確的偏好數據。男性公開宣稱的答案與私下選擇的答案之所以出現差距，我認為這是因為男性明白自己在公開場合存在外界壓力，知道自己應該怎麼回答才是合適的，不願意表露出真實的願望。在過去多年裡，公開答案與私下選擇之間的差距逐漸擴大。但如果不讓男性投票，而是給他們自由行動的空間，那麼這個差距就會消失。由此可見，男性的確處於一種可憐的境地。



圖1—4 「我就是喜歡高中女生。我變老了，她們的年齡永遠不變。」
同的結果。圖1—5中的灰色區域就是男性告訴我們的答案：

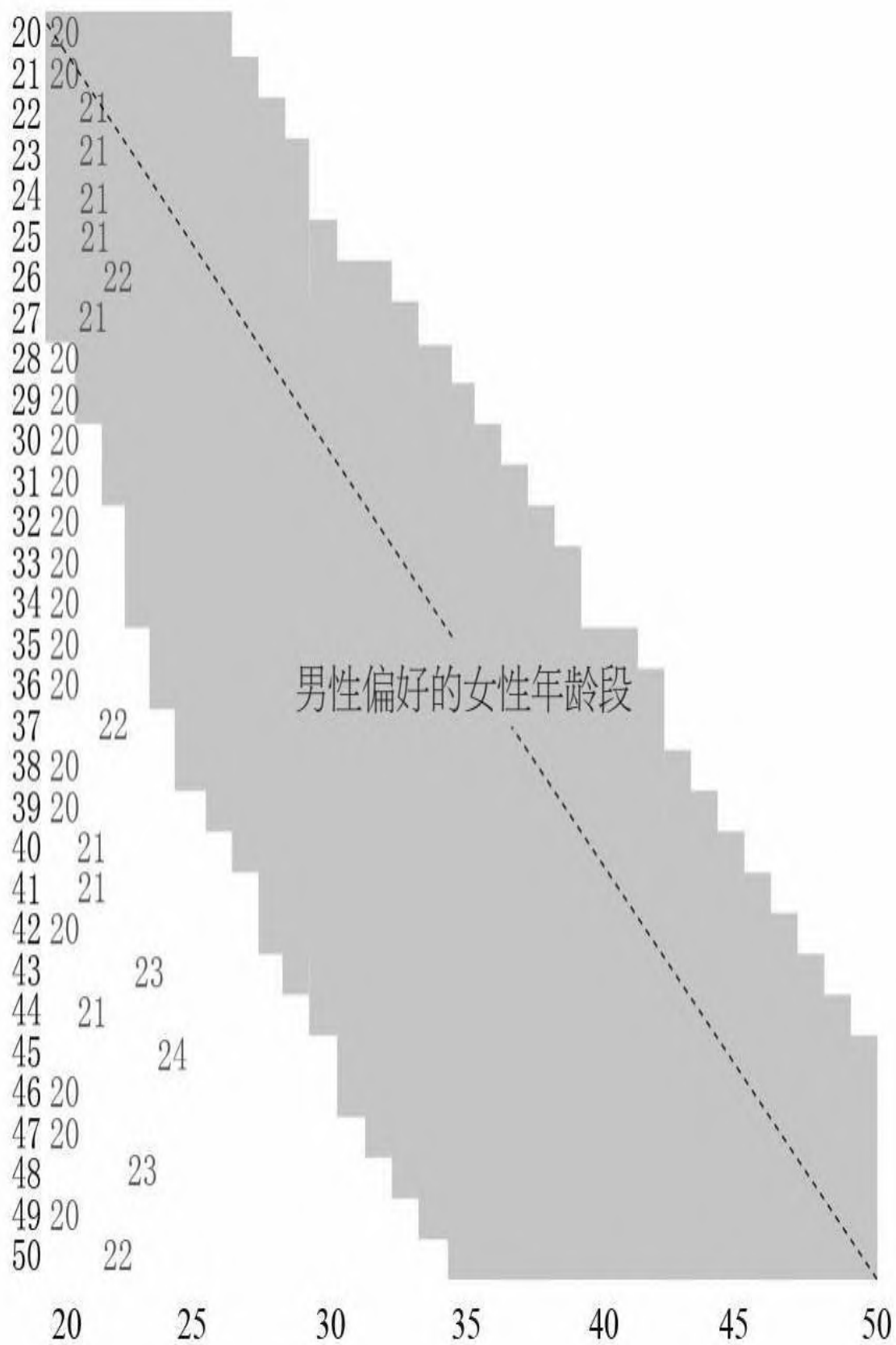


圖1—5 男性年齡vs男性聲稱的女性最有魅力的年齡

圖1—6（這類圖形中的最後一幅）揭示了男性最傾向於聯繫哪些女性。男性最願意聯繫的女性的年齡位於對角線左下側深灰色的區域。圖1—6下半部分中的三個垂直區域（30歲、35歲與40歲）揭示了男性走向中年時對女性認知的變化。你可以看到該圖中的明顯轉折。在44歲時，男性還試圖聯繫35歲或更年輕的女性，並且沒感到什麼不自在；而僅僅一年之後，即到了45歲時，男性則重新思考了一下對女性的態度。男性認為，如果與女性的年齡差距為9歲，一切還好；但如果相差10歲，則顯得太大了。

在現實生活中，男性與女性在建立浪漫關係的過程中，必須在「想要什麼」「說什麼」與「實際上怎麼做」之間實現平衡。無論男性在私下裡如何評價異性或青睞哪些女性，在現實中，50歲的男性成功追到20歲女孩的情況並不多見。一方面，社會傳統不支持這麼做；另一方面，約會需要雙方願意，一廂情願是行不通的。

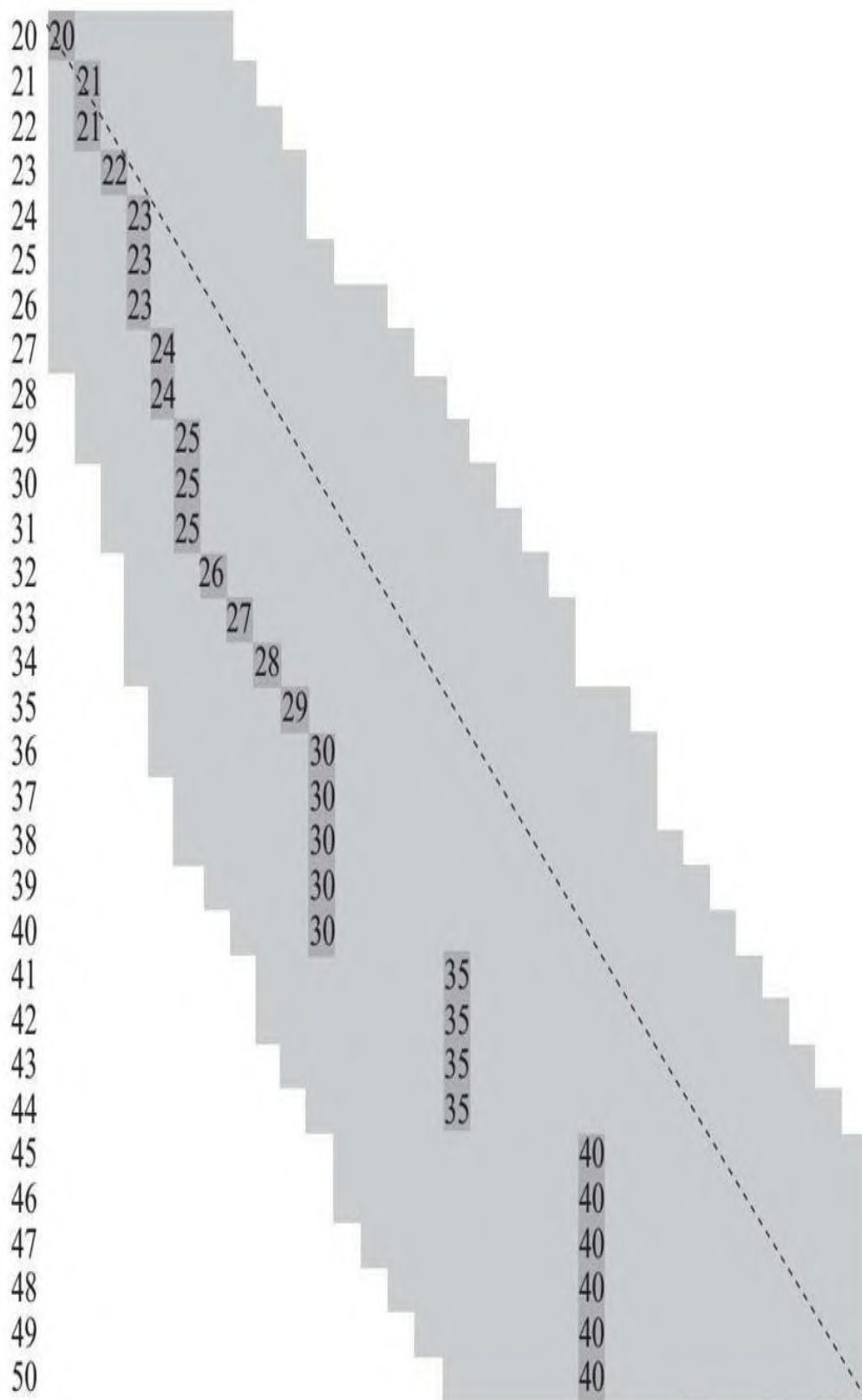


圖1—6 男性年齡vs男性最願意聯繫的女性的年齡

接下來我們分析一下女性主動聯繫男性的問題。我們在前面瞭解到，女性年齡的增加與女性眼中男性最有魅力的年齡要求呈現出了大抵一致的變化趨勢，再加上經濟狀況等一些非身體因素也促使女性接觸年齡較大的男性，所以，隨著男性年齡的增加，女性主動追求男性的次數會增加，而不是減少。這個趨勢一直維持到女性30歲出頭時。大概從這時起，女性主動追求男性的次數就會趨於減少，但女性失去男性追求者的速度更快。

想象一下這樣的情景：如果你是一名女性，和一個剛剛20歲的男性約會，由於他剛剛步入成年期，你可能會注意到對他感興趣的女性不止一名。如果你長時間關注他，就會發現他的追求者數量下降的主要原因是追求者移情別戀，結束了單身狀態。事實上，對他感興趣的女性可能會越來越多，因為隨著年齡漸長，他可能變得更富有、更成功，這些品質會吸引更年輕的女子。無論如何，男性的年齡都不是問題。在其頭20年的約會時間裡，雖然他和他的追求者都日益成熟，但那些女性仍然如同雙方都是20歲時那樣認為他是有魅力的。

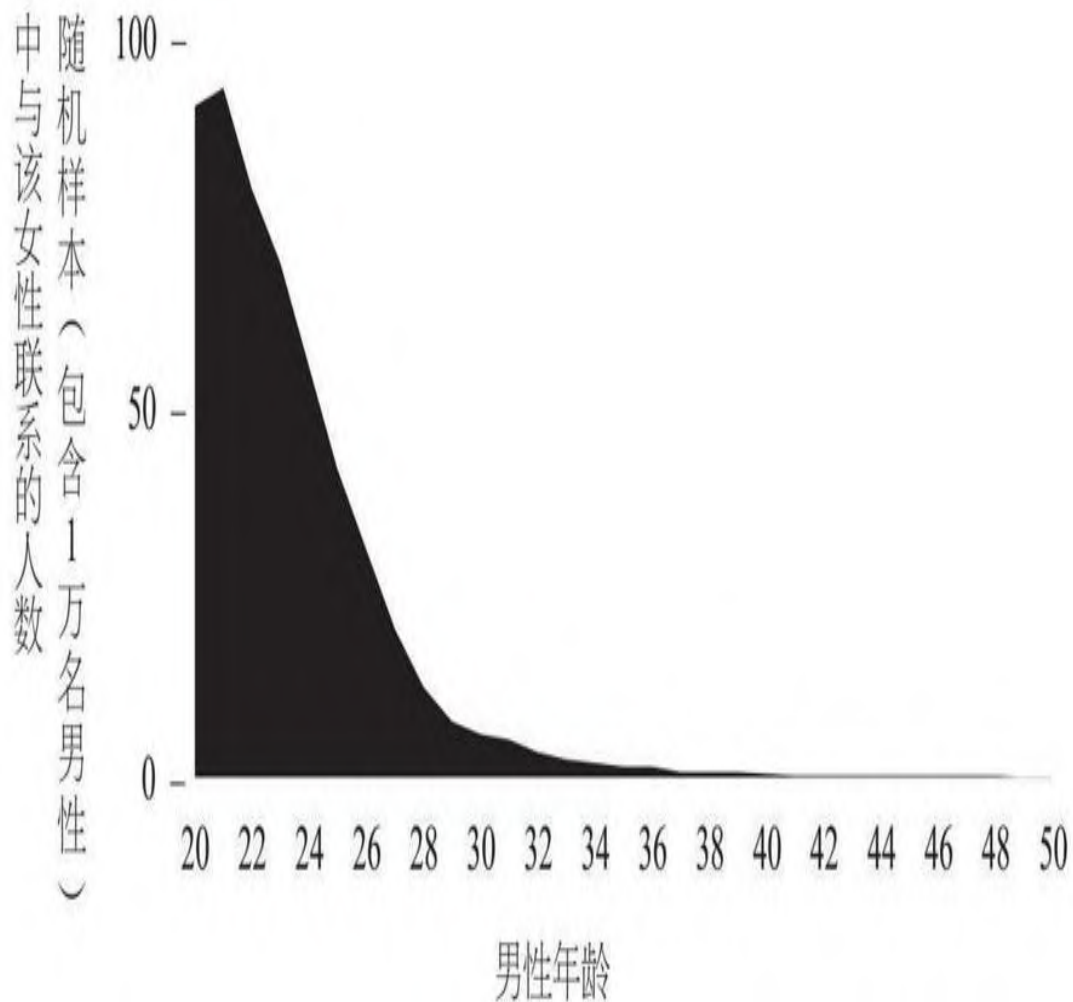
如果你是一名男性，和一個剛剛20歲的女性約會，那麼你可能遇到完全不同的情況。隨著年齡增長，她可能也會由於男性追求者結婚而失去一些追求者，但導致其失去男性追求者的主要原因在年齡。隨著年齡漸增，認為她仍有魅力的單身男性會越來越少。她的約會對象就像一個兩頭都有漏洞的罐子一樣，會加速減少。

從年齡來分析，單身男性的數量迅速減少。根據美國人口普查的結果，20~24歲的男性中有1000萬單身者，30~34歲的男性中只有500萬單身者，而40~44歲的男性中只有350萬單身者。^[3]如果根據女性對男性魅力的偏好來解釋這個下降趨勢，就會明白為什麼單身女性的選擇空間會越來越小。對於20歲的女性而言，如同圖1—7中的陰影部分所表示的那樣，其約會對象非常多。

她的同齡人（20歲出頭的男性）是與她聯繫最多的一個群體，隨著男性年齡漸增，與她聯繫的男性數量迅速減少，比如，30歲的男性聯繫人就很少了。雖然30歲的男性私下裡對20歲的女性表達了興趣，但他們不大可能花費很多精力去聯繫她們。此外，很多男性到30歲時都有了伴侶。當一名女性50歲時，如同圖1-8中所表示的那樣，對她感興趣的男性數量就微乎其微了。這時，她可能真的成了影片《BJ單身日記》

（*Bridget Jones's Diary*）中的大齡剩女布里奇特·瓊斯（Bridget

Jones)。



如果我們將圖1—7和圖1—8放在一起進行對比研究，就會發現，在對20歲女性感興趣的100名男性中間，只有9人對50歲的女性仍然有興趣。圖1—9與前面兩幅圖很相似，只是內容更加廣泛，揭示了女性25~50歲的追求者數量的變化。

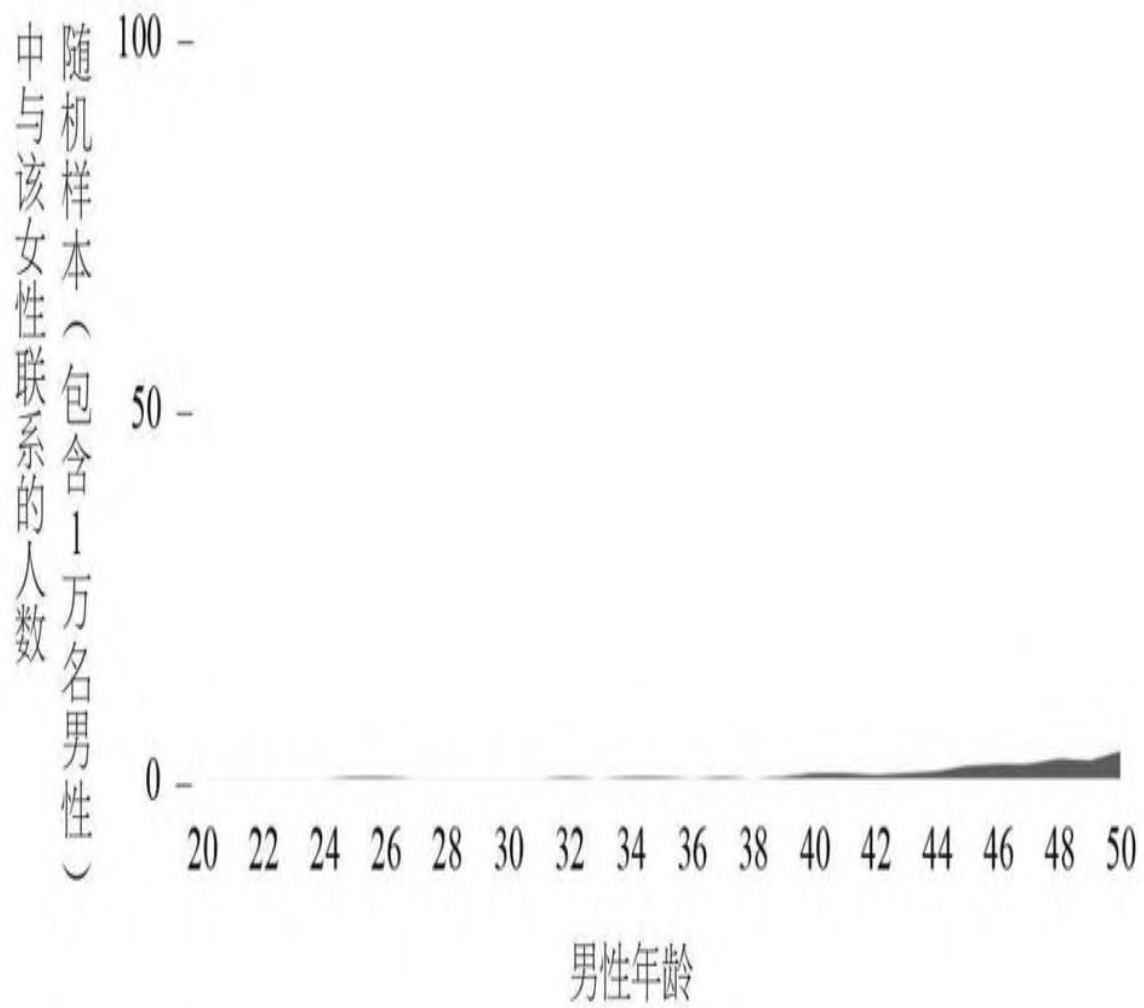


圖1—8 對一名50歲女性感興趣的男性的數量（按20~50歲的年齡來劃分）

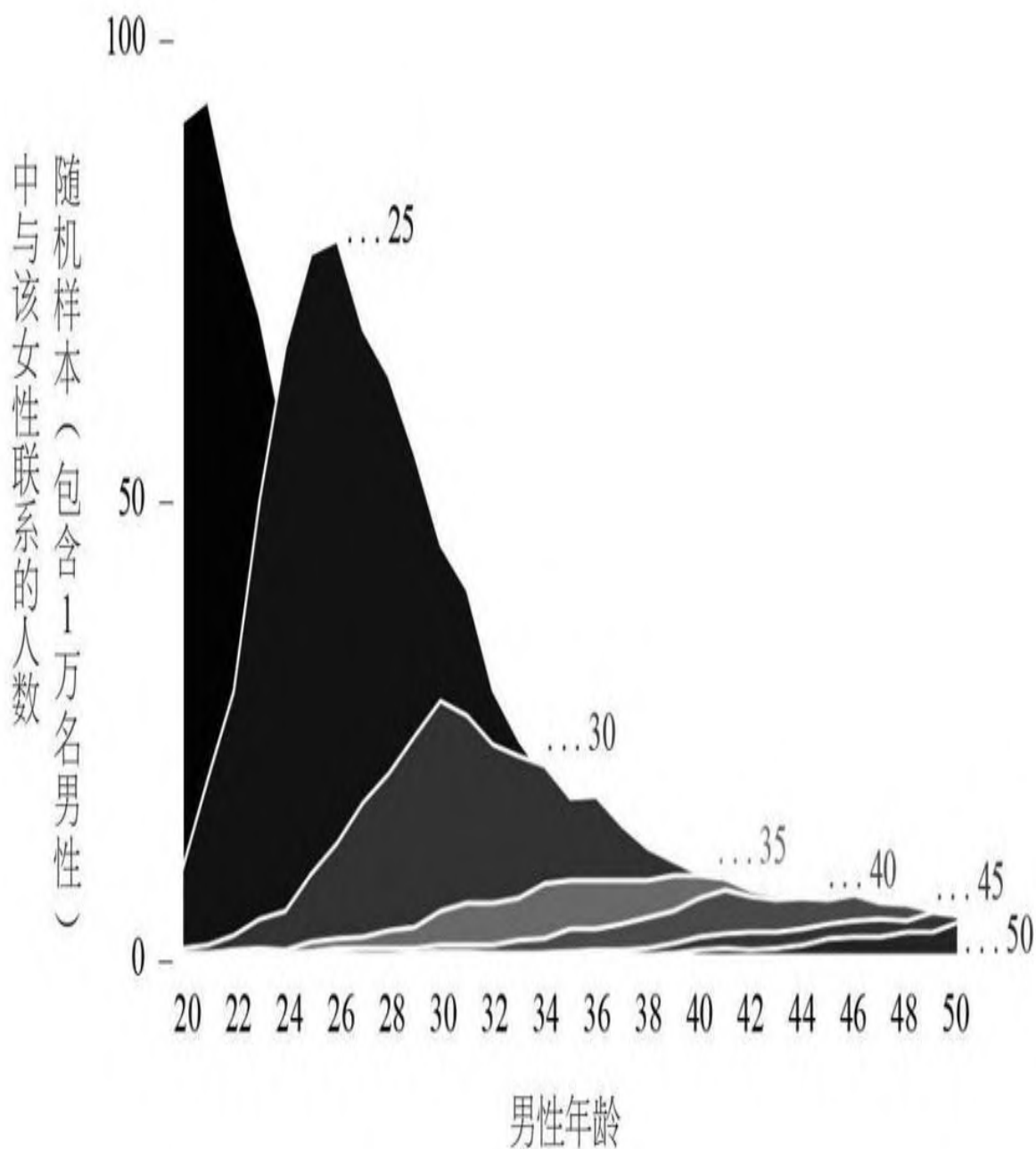


圖1—9 女性25~50歲時的追求者

在工作中，我經常看到兩個單身的人不知道由於什麼原因，就是不願意聯繫對方。這兩類人搜尋對方的時候總是會發生誤解：女性想要男性同她們一起變老，而男性永遠只青睞年輕女性。一位32歲的女性用戶註冊之後，將其偏好年齡設定在了28~35歲，過濾掉了一大批男性之後開始瀏覽。一名35歲的男性過來了，將其偏好年齡設定在了24~40歲，但幾乎不會聯繫超過29歲的女性。結果他們二人都沒有找到自己想要的

人。你可能會說他們就像黑夜中的兩艘船，但這樣說似乎也不完全準確。在我看來，男性的確像是在海上航行的船，駛向了大海深處，地平線漸行漸遠，而女性卻一直站在陸地上，沒有駛向海洋，任憑男性逐漸消失在視線外。

[1] 關於這一點，我最好解釋一下。我在將某個用戶的吸引力評分納入分析範圍之前，會確認其已經得到了至少25人給出的評分。因為「吸引力」評判起來的確存在很大的不確定性，所以，我認為如果評價人數少於25，算出來的平均評分可能不可靠。

[2] 我認為幾何結構使得數據可視化有別於一般的圖形和帶有數字的印象派圖表。就數據可視化而言，空間本身就可以傳遞出各個數據之間的關係。

[3] 數據源自2011年美國人口普查局發佈的調查結果，名為Marital Status of People 15 Years and Over, by Age, Sex, Personal Earnings, Race, and Hispanic Origin。

第二章 出醜效應

2002年，美國電影藝術與科學學院，即奧斯卡金像獎的主辦單位，請導演埃羅爾·莫里斯執導一部短片，講述我們為什麼熱愛電影。該學院希望這部短片以一組快速剪輯的人物鏡頭開場，這些人物中既有名人，也有普通人，每個人都講一講自己最喜歡的電影。我的朋友賈斯汀當時正在幫莫里斯導演挑選演員，便把我算上了。雖然不能保證最後的成片裡會有我，但我可以在鏡頭前接受採訪，看看短片的拍攝過程。

可能我受到了格外優待，和我同一天出境的都是些鼎鼎有名的人物：唐納德·特朗普、沃爾特·克朗凱特、伊基·波普、艾爾·夏普頓和米哈伊爾·戈爾巴喬夫。特朗普和戈爾巴喬夫當時是背對背的，但後來不知從哪兒就冒出了一張他們倆的照片，而且我還在中間。我在搶鏡流行起來之前就已經搶到鏡了，也蠻厲害的。之所以說「不知從哪兒」是因為賈斯汀本想拍一張特朗普的照片，但閃光燈剛滅，只聽特朗普打了個響指，保鏢就上來拿走了賈斯汀的相機。關於自己最喜歡的電影，特朗普選了《金剛》，他沒理由不喜歡一隻想要「征服紐約」的大猩猩。戈爾巴喬夫則通過一個大鬍子翻譯告訴我們，他最喜歡的電影是《角鬥士》。在短片第兩分零一秒出現的那個睜大眼睛、說最喜歡《天魔》的人就是我。

當時，我選擇的《天魔》是一部優秀的反基督教類電影。我之所以選擇這一部，或多或少有些隨機因素在裡面。現在，我仍然比大多數人都喜歡這類電影。其實，好電影太多了，我也不確定自己最喜歡哪一部。但至於自己最不喜歡哪一部，我是非常清楚的，那就是約翰·沃特斯執導、1998年上映的《派克》。這部電影我總共看過兩次，但每次都忍不住中途離場。第一次是和一群朋友去看的，但我實在受不了其混亂不堪的氛圍，更別提那誇張的口音了，只看了一半便不得不離開。第二個週末，又有一些朋友要去看這部電影。既然約翰·沃特斯是位備受尊崇的導演——嘿，關於這點我這樣的酷小夥當然知道——我想也許有可能是我第一次觀看的時候弄錯了呢，再說我也沒什麼別的事可幹，於是我便去看了第二次。

這種舉動是我22歲時才會有的短暫瘋狂。我並不是說約翰·沃特斯

的電影客觀上來看都很拙劣，只是它們並不適合我和其他很多人的口味，而沃特斯對這一點欣然接受，人們對他電影的排斥幾乎已經成為他作為導演的名片。這麼說吧，人們看完《派克》離開影院時，不會覺得它「乏味無趣」，你要麼愛得要死，要麼像我一樣看不到20分鐘就趕緊走人——還走了兩次。這種效果或許是導演本人有意為之的。

沃特斯的影迷們似乎因為同好之少而更加喜愛他。在OkCupid網站的用戶信息中搜索他的名字，得到的結果比搜索喬治·盧卡斯和斯蒂芬·斯皮爾伯格的結果加起來還要多。在Reddit上，他甚至有自己的專屬子頁/r/JohnWaters。^[1]雖然網頁的瀏覽量不是最大，但影迷們卻真的在上面放置了新聞、老視頻片段、關於他的問題、評論等資源。網站上也有喬治·盧卡斯的子頁/r/GeorgeLucas，但上面直到現在也僅有一條留言；而如果你在地址欄裡輸入/r/StevenSpielberg進行搜索，Reddit的服務器會告訴你「未找到相關信息」。因為雖然斯皮爾伯格的作品很棒，但沒有人會有足夠的熱情為他創建一個網頁。甚至像J.J.艾布拉姆斯（J.J.Abrams）這樣在互聯網上非常活躍的導演都沒有自己的個人網頁。如果你要粉絲為你創建一個網站，那麼他們就得具備特別強大的動力，而如果你是一個與眾不同的人，並遭到了眾人的圍攻，那麼你的粉絲往往具有更強大的動力。全情投入的熱愛就好比活塞裡的蒸汽，壓力更能使其有力迸發。

就像他之前及之後的許多藝術家一樣，沃特斯十分清楚一個道理：一些人對你的排斥會將另一些人變得跟你更為緊密。我提起沃特斯並非僅僅因為《派克》一直讓我很糾結，更主要是因為沃特斯深知這一道理具有普遍適用性，知道它並非僅適用於藝術。他說過很多非常有道理的話，但有一句直擊我心底：「你永遠也不能忘懷的臉龐便是美的。人的面貌應該張揚有特點，而非平淡無奇。」^[2]他說得完全正確，不管是對於音樂、電影還是各種各樣的人類現象來說都是如此，因為缺陷會給人帶來強有力的震撼。即使在人際關係領域也是如此，如果大家都喜歡你，那就意味著你會受到相對的忽視；如果有人不喜歡你，那麼一些人對你的喜歡程度則會更深。放到具體情形來看，對於一名女性而言，如果有的男性覺得她相貌醜陋的話，她總體上的性魅力反而會提升。

這一點在OkCupid網站的信息評級上可以清楚地看出來。該網站的評級系統是5星級，所需提交的評級也比簡單的「是」或「否」要更深入。用戶提交不同星級的評價，也給我們創造了探索的空間。為了印證這項發現，我們必須踏上一段短短的數學旅程。正是這些小小的鍛鍊使

得數據科學發揮了巨大作用。想要完成一幅拼圖，首先得把所有的圖塊擺開，然後開始嘗試拼出圖案。如果沒有經過仔細的篩選、刪減和慎重使用，那就基本上不可能從海量原始數據中尋得一絲靈光。

接下來，我們考慮一組魅力相近的女性，比如，圖2—1中獲得中等評分的女性。

我們以這組中的某一名女性為例，考慮一下男性可能給她的評分，進而分析一下為什麼她的得分會處於中等水平。男性的評分有數千種可能，下面只是我設想的5個評分模式，分別為1分、2分、3分、4分和5分，最終平均分為3分（見表2—1）。

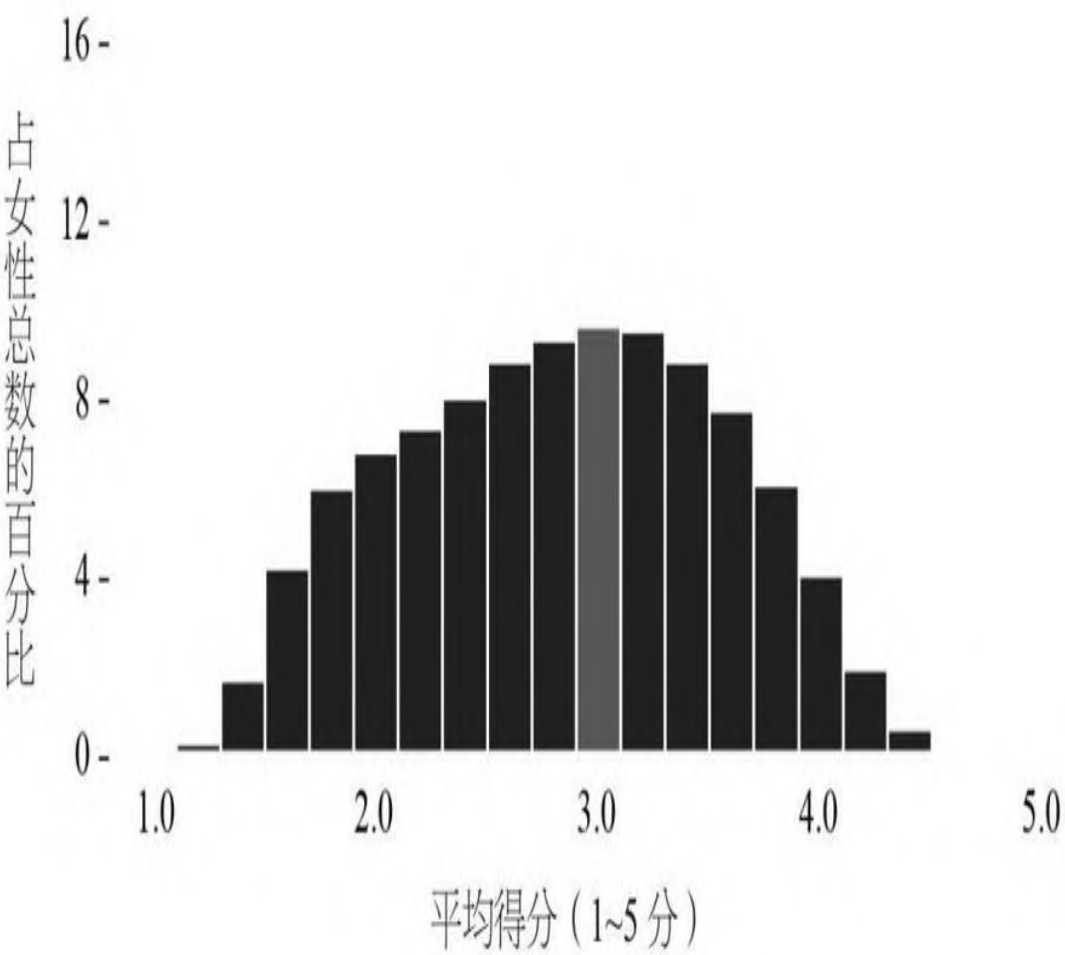


圖2—1 受到男性評價的女性
表2—1 參與評分的男性人數

	1	2	3	4	5	平均得分
模式A			100			3.0
模式B		10	80	10		3.0
模式C	10	20	40	20	10	3.0
模式D	25	25		25	25	3.0
模式E	50				20	3.0

你可能已經注意到了，從模式A到模式E，男性給出的評分呈現出越來越明顯的極端化傾向。雖然每一種評分模式平均得分都是3分，但其表現方式卻呈現出極大的差異。模式A是一種共識的體現，表明參與投票的男性完全一致地認為這名女性的魅力正好屬於中等。然而，看一看模式E我們就會發現，雖然男性評分的平均值仍然是3分，但事實上沒有一名男性認為這名女性的魅力屬於中等。模式E代表了最極端的評分，也就是說，對於我們假想中的這名女性，每當一個男性評出1分，就有一名男性評出5分，最終結果仍然是3分。這跟約翰·沃特斯得到的評分情況具有一定的相似性。

這些評分模式體現出了「方差」這個數學概念。^[3] 方差用來度量各個隨機變量與其均値之間的偏離程度。方差越大，數據的波動越大；方差越小，數據的波動就越小。在表2—1中，方差最大的是模式E。方差概念應用得最為普遍的地方就是衡量金融市場上的波動性和風險。我們考慮一下圖2—2中的兩家企業：

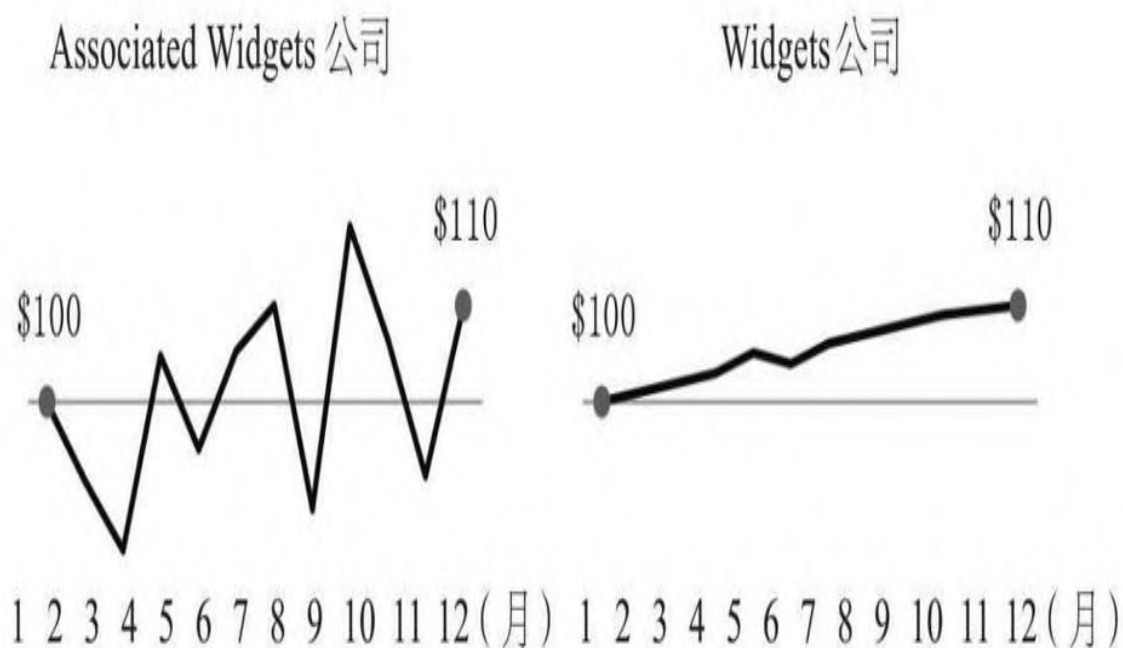


圖2—2 兩家公司的對比

兩家企業的年度利潤率都是10%，但它們代表的盈利模式卻大相逕庭。第一家企業的盈利波動性較大，而第二家的盈利則逐步增長，每個月都有盈利。通過計算方差，分析人士就能很明顯地發現這一趨勢。如果其他條件相同，投資者更願意選擇第二家企業，因為兩家企業的收益是相等的，但選第二家企業少一些擔心。當然，從浪漫的角度來講，擔心往往意味著收益，很多人之所以選擇高風險投資，關鍵就在這裡。方差不僅能解釋投資者的選擇，也能解釋男性在追求女性過程中的選擇。男性對一名女性魅力的評分，如果呈現出了較大的方差，那麼這名女性得到的關注就較多；如果方差較小，得到的關注則會較少。

在一組女性中，有的是選美皇后，有的是居家少婦，也有很多介於二者之間，她們看上去可能同樣漂亮，但她們得到的私信數量與男性對其評分的方差具有高度相關性。得到最多關注的人，肯定是方差較大的人。方差產生的影響不可小覷，因為如果男性對一名女性的評分分化得較為嚴重，也就是說方差較大，那麼這名女性得到的私信數量可能比其他女性多出70%。這就意味著男性評分波動足夠大的話，完全能夠讓一名女性在約會排行榜上上升好幾檔。比如，如果按照評分的均值來講，一名女性可能屬於均值最低的20%，另一名女性可能屬於均值最高的30%，而如果第一名女性獲得評分的方差較大，那麼她獲得的約會機會

與第二名女性獲得的約會機會可能幾乎是相同的。

之所以出現這種現象，部分原因就在於方差意味著有很多人非常喜歡你，同時也有很多人很不喜歡你。那些對你非常熱情的人，我們姑且稱之為「狂熱粉絲」，就是那些給你發私信最多的人。這些人很可能給你評出5分。有了這些人的支持，你獲得的私信數量自然會比較多。

但不要忘了，還有一些人會給你消極的評價，但這些消極評價反而能夠讓你得到更多的關注。比如，在表2—2中，C女和D女獲得的平均分雖然低於A女和B女，但獲得的約會機會卻比前兩類女性還多出10%左右。

表2—2 參與評分的女性人數

	1	2	3	4	5	平均得分
A女	2	22	27	29	20	3.4
B女	10	13	31	28	18	3.3
C女	32	22	12	16	18	2.7
D女	47	13	6	19	15	2.4

我一直討論私信的問題，似乎用戶發私信的目的就是單純地發私信，但在交友網站上，私信卻預示著很多結果。發了私信之後，用戶之間可能會開展深度對話，交換聯絡信息，並最終達成線下會面。評分方差越大的人，獲得這些機會的次數就越多。比如，與A女相比，雖然D女獲得的評分可能低很多，但D女與男性開展深度對話的次數可能多出10%，約會次數可能多出10%，而且性愛次數可能也會多出10%。

此外，如果一名男性給一名女性的評分是1分或2分，那麼這名男性幾乎不會與這名女性約會。事實上，如果一名男性給一名女性的評分非常低，那麼他幾乎不會聯繫她。^[4]但正是由於某些人不喜歡你，恰恰會

促使其他人更加喜歡你，讓你獲得更多關注。美國前總統小布什的智囊卡爾·羅夫深諳此道，他一方面將小布什打造成完美的人物，一方面又讓小布什偶爾流露出一些無傷大雅的缺陷，在招來一部分人討厭的同時，也讓小布什獲得了更多選民的青睞。

雖然我談到了方差具有如此重要的意義，但我的OkCupid網站從來不會發布關於任何用戶的原始評分數據，當然也不會發布這些評分數據呈現出的方差。這種情況似乎有些不可思議。其實，我們之所以不發佈這類數據，是因為用戶不會有意識地根據這些數據來做決策。不過，人們能有意或無意地感覺到這些數據背後的數學規律。比如，當一名男性被一名不太符合傳統審美標準的女性吸引時，他肯定會經歷以下心理活動。他會認為，這名女性的外貌不太符合傳統審美標準，這就意味著其他一些男性有可能不會追求她，這樣就減少了自己追求這名女性時面對的競爭；對手減少了，自己成功的可能性就大了。我可以想象我們的男性用戶瀏覽她的頭像時，鼠標的光標一直圍繞著她的頭像旋轉，他會暗自思忖：我打賭，在她遇到的人裡面，肯定不會有很多人認為她長得很美。其實，她的獨特性非但不會讓我生厭，反而讓我更加喜歡。她就是一塊未經雕琢的璞玉。從某種程度上來講，這名女性身上的一些不受歡迎的特徵，正是對這名男性產生吸引力的地方。如果這名男性正在猶豫是否要介紹自己，女性的這些特徵反而可能鼓勵他去表白。

我們也可以從相反的角度看待這一現象。對於一個比較有魅力、符合傳統審美標準的女性而言，男性的評分可能都比較高，也就是說，這些評分的方差較小。這名女性可能非常受歡迎。但正因為她受到了每一名男性的歡迎，可能會給男性留下這樣一個印象：追她的人肯定不少，這會增加我面臨的競爭力。這樣一來，那些原本感興趣、猶豫要不要表白的男性就將目光轉移到其他女性身上了，這名女性得到的約會機會反而減少了。

這就是我的理論，但其他許多領域也證明了方差會產生積極的影響。社會心理學家稱之為「出醜效應」^[5]。只要你是一個非常有能力的人，偶爾犯一個小錯誤，不僅瑕不掩瑜，反而更使人覺得你和別人一樣可能犯錯。這反而成為你的優點，讓人更加喜愛你。瑕疵會讓你的優點更加鮮明。「不完美」的事物也具有存在的必要性與合理性，我們人類大腦的運作機理就體現了這一點。作為與大腦的情感區域聯繫最為密切的一種感覺，我們的嗅覺就喜歡混雜的味道，而不是單一的味道。^[6]科學家們在實驗室裡把少量的臭味與芳香混合在一起，結果發現這樣會增

強芳香，對人類嗅覺帶來更大的刺激。但在科學家們進行論證很久之前，人類器官在漫長的自然進化過程中就已經意識到這一點了。橙花和茉莉花等鮮花的芳香中就包含一個重要的成分——吲哚。吲哚在蛋白質中所佔比例約為3%。這種物質在大腸裡面是很常見的。就其本身而言，它具有臭味，但如果沒有吲哚的存在，花兒的芳香就會打折扣。吲哚也是人工合成香水中的一種原料。

模特世界中也體現出了OkCupid網站數據揭示的「出醜效應」。這些模特肯定都是非常靚麗的，當然也能獲得滿分5分的評價。但即便對於這些人而言，也需要通過一些無傷大雅的瑕疵來將自己同其他人區別開來。比如，超級名模辛迪·克勞馥（Cindy Crawford）長得十分清秀，美中不足的是嘴角有一顆非常扎眼的黑痣，但她卻覺得這顆痣使她更顯嫵媚，使她的美更有特色。嘴角有痣的她，蘊含著一種桀驁不馴、個性四射的美，迅速被大眾接受，事業開始騰飛，媒體稱之為「最有性格的一位典型美國美人」。超模琳達·伊萬格麗斯塔（Linda Evangelista）的髮型的確不太符合傳統的審美標準，你絕對不會說她的髮型使她顯得更靚麗，但正是得益於這種非同尋常的髮型，她在眾人眼中才變得更有氣質。從模特行業的體重標準來看，凱特·厄普頓（Kate Upton）的體重似乎多了幾磅，但這反而為其平添了幾分魅力。如果只分析那些身著泳裝的模特，似乎沒有代表性，接下來，我根據自己收集到的數據舉幾個關於普通人的例子，看看「出醜效應」在普通人身上的體現。圖2—3中的6名女性獲得的平均得分都屬於中等，但她們往往引起男性給出一些極端的評分：很多男性的評分非常高，另一些男性的評分卻非常低，而介於二者中間的普通評分卻非常少。^[7]

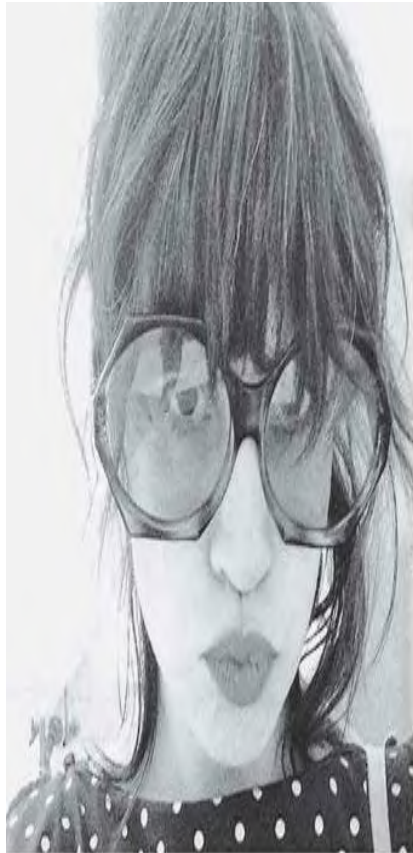


圖2—3 平均得分中等的6名女性

非常感謝她們自信地同意我把她們作為案例進行展示和討論。她們很有代表性。這些人在上傳頭像時故意選擇了一些非同尋常的情景，有的在展示人體藝術，有的做出了嘲諷的表情，有的在吃奶酪，有的還露出了壞壞的面部表情。你會發現許多相對正常的女性都有一個非同尋常的特徵，比如下排中間的那名女性把秀髮染成了藍色，只是本書給出的是黑白照片，你看不出來而已。如果一名女性刻意突出自己的某一項優點或缺點，那麼她就會顯得更加出眾，你就能更加容易注意到她。比如，如果你體重偏重或有文身，卻將其作為自己的一個特點突出表現出來，那麼雖然很多人不喜歡，你獲得的評分均值仍然有可能達到3.3的水平，這樣你反而會產生更大的影響力。

因此，歸根結底，因為每個人都有這樣或那樣的缺陷，我們就要堅持這樣一條真理：勇敢面對缺陷，做真實的自己。我相信，為了迎合他人而刻意去掩飾自己是非常危險的做法，無異於給自己套上了一層又一層的面具，這是違反科學的，只會適得其反。很多父母在給孩子提建議時，就會注意到這一點。比如，如果一個大鼻子、戴牙套的14歲男孩不明白自己為什麼不太受歡迎，那麼很多父母提出的建議就是坦然面對。這類建議符合我們的數據揭示出來的「出醜效應」。如同我在前面所說的那樣，很多人都能感受到日常事物背後的數學規律，幸運的是，母親們能更加清楚地感受到這一點。我多麼希望我的母親在我九年級時告訴我天才其實並不是完美的。



[1] 在Reddit網站上，這些子頁被稱為subreddit，我在後面會更加清楚地解釋這個網站的內容。

[2] 引自約翰·沃特斯（John Waters）於2005年出版的《衝擊價值》（Shock Value: A Tasteful Book About Bad Taste）一書第128頁。

[3] 本章提到的「方差」，都用標準方差來衡量。

[4] 在OkCupid網站上，只有0.2%的用戶會向自己認為評分應該低於3分的人發送私信。

[5] 只要在谷歌搜索一下「出醜效應」，就能發現很多例子。丹尼特·費因—加爾、扎卡里·托馬拉和希夫·托馬拉（Danit Ein-Gar, Zakary Tormala, and Shiv Tormala）於2012年在《消費者研究期刊》（Journal of Consumer Research）第5期上發表了《當缺陷通向綻放：消極信息的積極影響》（When Blemishing Leads to Blossoming: The Positive Effect of Negative Information）一文，比爾·斯奈德（Bill Snyder）針對此文寫了一篇名為《消極信息的積極影響》（The Positive Effect of Negative Information）的摘要，我著重參考了這篇摘要。

[6] 關於這一段，我著重參考了費邊·格拉本赫斯特（Fabian Grabenhorst）等人於2007年在《神經科學期刊》（Journal of Neuroscience）上發表的《令人愉悅和不悅的刺激因素如何在不同大腦區的域組合方式：基於氣味混合物的研究》（How Pleasant and Unpleasant Stimuli Combine in Different Brain Regions: Odor Mixtures）一文。維基百科的「吶哖」條目也描述了它有強烈的臭味。如需進一步瞭解吶哖在香水和天然花香中起到的作用，請像我所做的那樣，參考下面這個鏈接中的文章：perfumeshrine.blogspot.com/2010/05/jasmine-indolic-vs-non-indolic.html。

[7] 為了保護用戶隱私，我在徵求用戶許可的時候採取了雙盲系統。具體來講，我先把各項指標（比如，女性、所獲評分的方差、整體吸引力居於中間水平）發給OkCupid的數據分析團隊。該團隊根據我的各項指標自動生成一個潛在的名單（只有名字，而沒有照片和其他相關資料），然後將這個名單發送給管理員，由管理員聯繫名單上的用戶，詢問這些用戶是否同意我使用她們的照片。（經常有媒體向我們詢問是否可以使用我們用戶的照片，我們也經常徵求用戶的意見，因此，這種事情並不鮮見。）只有用戶允許我們使用之後，我們才能獲得用戶的照片和其他相關資料。

第三章 「作家」的世界

思鄉情結一度被稱為「瑞士病」^[1]，因為瑞士的僱傭兵曾經遍佈歐洲，卻因為希望回家而臭名昭著。有時候，他們不願意參加戰鬥，而是淚眼矇矓地唱起了牧民的歌謠。如果你是法國國王，僱了這些瑞士軍人同胡格諾派教徒戰鬥，他們的歌謠是無濟於事的。後來，這些歌謠就被禁止了。在美國南北戰爭期間，思鄉情緒一度也是一個非常嚴重的問題。大約5 000名士兵因為思念家鄉而喪失了戰鬥力，74名士兵因為思鄉心切而患病去世，至少軍醫的醫療記錄是這麼寫的。鑑於當時的情形，因過度悲傷而病亡是可以理解的，但話又說回來，當時也是水蛭等害蟲氾濫成災的時代，誰也無法確切地知道這些士兵究竟是死於疾病，還是死於思鄉。但當時很多離開家鄉走向戰場的人的確都具有濃重的思鄉情結，很多早期的文學作品都提到了士兵的思鄉情結，將這種情結視為一種真正意義上的疾病。1863年，為了緩解士兵們的思鄉情緒，美國科學家們在波托馬克河畔爭分奪秒地開發一項終極武器，即「高中年鑑」。每當我的思緒回到那段黯淡的內戰歲月，總是忍不住想象他們製作年鑑的情景。

事實上，我不知道現在是否還有高中年鑑。既然現在有了Facebook，人們往往難以理解為什麼還要高中年鑑。Facebook曾經在季度報告中表示，未滿18歲者使用Facebook的數量呈現出了減少趨勢。^[2]孩子們是不是迴歸紙質交流方式了呢？我也不知道。^[3]但無論青少年用哪種方式保持聯繫（用Snapchat、WhatsApp或Twitter），我相信他們不是通過語音交流的，而是通過文字交流。上述這幾個網站的服務中，分享照片顯然是一項非常有吸引力的服務，但如果沒有鍵盤，你只能傳一張照片。即使在Instagram上，照片畢竟只有幾平方英寸（1平方英寸=6.45平方釐米），評論和標題還是很有必要的。但是，文字畢竟是文字，我們今天仍然主要依靠文字來表達自己的感覺，並與他人建立聯繫。

事實上，雖然技術對我們文化的影響令我感到絕望，但我敢肯定，與20世紀90年代的我和我的同學們相比，即使2014年最沉默的少年寫出的文字也算比較多的。因為我們那個年代沒有今天這些以文字為基礎的

交流媒介，如果你需要和其他人交流，打電話是絕大多數人的選擇，一年到頭可能只寫一些「謝謝你」之類的便條，也可能會寫一封信，從字數來看，肯定還沒有今天一名普通高中生在一個上午打出來的字數多。在我看來，互聯網固然有很多令人遺憾的負面影響，但具有一個無可取代的優勢：互聯網世界是一個「作家」的世界。互聯網時代讓每一個人都成了「作家」。你的在線生活都是通過打字實現的，只要你會打字，就能在互聯網上工作、社交或調情。我真切地感覺到互聯網時代的文字交流就像奧斯汀時代的書信交流那樣偉大。無論我們使用的是什麼文字，無論我們的打字方式如何，我們彼此相互交流的頻率都比以前有了大幅提升。

∞

在美國內戰期間，蘇利文·巴魯（Sullivan Ballou）少校是北方聯邦的一員，駐紮在波托馬克河畔，飽受思鄉之苦。1861年7月14日，在南北戰爭期間第一次大規模戰役——馬納薩斯戰役爆發前夕，他給妻子「最親愛的莎拉」寫了一封告別信。一個星期之後，他就在這次戰役中陣亡了。這是他給家人的最後一封信，這封信描述了這個國家在未來幾年裡面臨的悲哀與傷痛。在肯·伯恩斯（Ken Burns）執導的紀錄片《南北戰爭》中，就有一名解說員聲情並茂地讀出了這封信，這個情節在這部紀錄片中具有重要的意義。巴魯流露的真情和對國家的忠誠打動了千萬人的心靈，而且這封信被曝光並得到了傳播，堪稱人類有史以來最有名的信函之一。當我在谷歌上搜索「著名的信」（famous letter）時，谷歌列出的第二個選項就是這封信。這是一件美好的作品，但除了這封信函之外，肯定還有其他美好的作品，只是被燒掉了、在混亂中丟失了、被風吹走了，或者逐漸腐爛了，以致沒有流傳下來讓我們有機會讀到。

之前的作品如果保存得好，就能流傳下來，否則就會失傳。而到了今天的互聯網時代，由於一切都能得到很好的保存，我們再也不必靠運氣來了解先人的想法或話語了，也不必讓一個人來代表其他人。我們個人努力過程的前前後後，甚至過程的細節，都得到了很好的記錄。現在，你可以在視頻網站YouTube上找到關於蘇利文·巴魯那封信的各種音頻和視頻，瞭解到不同人的解讀方式。^[4]很多評論都說，「很多解讀都讓這封信變了味兒」。的確如此。但他們——或者說「我們」的做法創造了一種別樣的豐富多彩，一種別樣的美好，我們譜寫的不是抒情辭藻組成的詩篇，而是不同思維、不同解讀相互並存的詩篇。這些不同解讀

方式的並存反映出人類交流方式正在發生重大的改變。這些新的交流方式催生了社區意識，密切了人與人之間的聯繫。

如果你想了解人們的寫作方式，那麼最佳的著眼點就是研究一下他們在毫無防備狀態下寫出來的未經雕琢的詞語。我們有很多這類數據。未來兩年內，人們在Twitter上打的字比人類有史以來印刷的書籍包含的字數還要多。^[5]這是新型交流方式的縮影。這種新型交流方式的兩大特點就是簡短和即時。事實上，Twitter是第一個鼓勵簡短和即時的網站，而且是第一個對這兩大特徵提出要求的網站。Twitter用戶在發帖時，會受到140個字符的限制。為什麼要限制140個字符呢？就是為了讓你在140個字符內把一件事說清楚，這樣可以讓文字精簡，讓世界更容易瞭解「正在發生的事」。Twitter突然流行起來，突然重新定義了我們的寫作方式，似乎證實了人們對互聯網正在「扼殺我們的文化」的恐懼。在這種備受限制的空間裡，人們怎麼繼續好好寫東西呢？怎麼繼續好好思考呢？頭腦受到如此約束之後，會變成什麼樣子呢？演員拉爾夫·費因斯（Ralph Fiennes）的一句話道出了很多人的心聲：「你只需要看一看Twitter，就能找到大量證據來佐證這一事實，即莎士比亞的戲劇或沃德豪斯（P.G.Wodehouse）的小說裡所用的英語單詞，現在用得很少了，人們甚至不知道它們的意思了。」^[6]

但即便稍微做一個基本的分析，我們就會發現Twitter上的語言其實並沒有退化。^[7]接下來，我從Twitter和牛津英語語料庫（OEC）中找出了一些最常用的詞，對它們進行了對比分析。牛津英語語料庫中的詞條已經將近25億條，其中不僅包括單詞，還包括短語、句子以及用法和拼寫示例等，來源包括新聞、小說、博客、報紙等，是對當代英語詞彙的規範性普查。^[8]人們最常用的單詞多達好幾萬個，但我只列出了100個，這個樣本雖然看似微不足道，但幾乎任何文章的寫作都離不開這些單詞（Twitter和牛津英語語料庫中都是如此）。首先，在Twitter這一欄，我們要注意這樣一個非常重要的事實：雖然很多人都在抱怨網絡英語已經退化了，但在這100個最常用的單詞中，只有兩個真正意義上的網絡用語，即rt代表retweet（轉發），u代表you（你，你們）。你可能認為縮略語是Twitter字數限制的產物，但人們會想辦法繞開這種限制，比如一篇長文分成幾個部分發出，這樣每個部分都不會超出字數限制。第二個要注意的事實就是，如果你計算一下兩欄中的單詞的長度，你就會發現Twitter那一欄的平均長度是4.3個字母，而語料庫那一欄的平均長度是3.4個字母。且不看單詞長度，我們從單詞的內容角度做個對比分析。為了便於對比，我把Twitter那一欄的幾個特有單詞標成了灰色（見

表3—1)。

OEC一欄的單詞看起來有些單調乏味，有很多輔助性和修飾性的詞語幫助你理解有關名詞或動詞，而在Twitter上沒有這類詞語的空間，因為每個詞語都是獨立運用的，沒必要用這些單調乏味的詞語去解釋另外一些詞語。所以，你就看到love（愛）、happy（幸福）、life（生活）、today（今天）、best（最佳）、never（從不）、home（家）等具有感情內涵和生活氣息的詞語出現在了最常用的100個單詞裡面。事實上，由於Twitter迫使用戶用較少的單詞來表達完整的意思，所以可能會提高他們的寫作水平。這體現了小威廉·斯特倫克（William Strunk Jr.）的著名格言——「省略不必要的詞」。Twitter用戶除了追求言簡意賅之外別無選擇，而且字數限制其實解釋了為什麼Twitter那一欄單詞的平均長度長於OEC那一欄的單詞，因為在字數有限的情況下，較長的單詞意味著它們之間的間距較小，這意味著空間的浪費也較小。雖然Twitter用戶表達思想的長度可能縮短了，但沒有證據表明思想的深度有所下降。

表3—1 Twitter與OEC100個單詞對比分析

	OEC	Twitter		OEC	Twitter
1	the	to	51	when	back
	be	a		make	an
	to	i		can	see
	of	the		like	more
	and	and		time	by
	a	in		no	today
	in	you		just	twitter
	that	my		him	or
	have	for		know	as
10	I	on	60	take	make
	it	of		people	who
	for	it		into	got
	not	me		year	here
	on	this		your	want
	with	with		good	need
	he	at		some	happy
	as	just		could	too
	you	so		them	u
	do	be		see	best
20	at	rt	70	other	people
	this	out		than	some
	but	that		then	they
	his	have		now	life
	by	your		look	there
	from	all		only	think
	they	up		come	going
	we	love		its	why
	say	do		over	he
	her	what		think	really
30	she	like	80	also	way
	or	not		back	come
	an	get		after	much
	will	no		use	only
	my	good		two	off
	one	but		how	still
	all	new		our	right
	would	can		work	night
	there	if		first	home
	their	day		well	say
40	what	now	90	way	great
	so	time		even	never
	up	from		new	work
	out	go		want	would
	if	how		because	last
	about	we		any	first
	who	will		these	over
	get	one		give	take
	which	about		day	its
	go	know		most	better
50	me	when	100	us	them

馬克·利伯曼（Mark Liberman）^[9]是賓夕法尼亞大學的一位語言學教授。他也總結出了同樣的現象，直接呼應了費因斯先生的研究成果。根據他的計算，《哈姆雷特》中使用單詞平均包括3.99個字母，英國小說家沃德豪斯的故事採用的單詞平均包括4.05個字母，而Twitter上的單詞平均包括4.8個字母。^[10]像他這樣從Twitter上挖掘數據的比較語言學家還有很多。亞利桑那州的一個研究團隊已經超越了對比字數和長度的層次，而開始對人們的寫作風格和情感類型進行對比研究。^[11]他們得出了一些令人驚訝的結論。第一，他們發現Twitter並沒有改變一個人的寫作方式。他們跟蹤了很多人的舉了很多例子來佐證自己的觀點，其中一個例子就是如果一個作家在電子郵件或短信中使用「u」來表達第二人稱，那麼無論是在Twitter還是在其他情況下，這位作家仍然會保持這一風格。相似地，如果這位作家一般採用you的寫作方式，那麼他在Twitter、寫電子郵件、發短信或其他情況下也會這樣寫。至於第一人稱單數是採用I還是i，也具有高度一致性。也就是說，一個人的寫作風格不會因為交流媒介的變化而改變，也不會出現簡化。你原本如何寫的，就會一直堅持下去，不管在哪兒寫都是如此。語言學家們還測量了Twitter的單詞密度以及具有真實意義的單詞（如動詞和名詞）的比例，結果發現其比例不僅高於電子郵件，甚至還高於政治評論網站Slate上的文字。一切都指向了相同的結論：Twitter的字數限制並沒有像人們所想象的那樣改變了我們的寫作方式。通過分析我收集的數據，我們發現的不是一片只剩下樹樁的荒地，而是一片茂密的樹林。

這種深入的分析（單詞密度、詞頻）暗示了寫作方式轉換的真實性質。Twitter給語言研究領域帶來的變化遠遠超出了給語言本身帶來的變化。Twitter讓我們認識到，語言不僅是組織思維的基礎構件，還是人類社會的黏合劑，這才是語言的原始意義。與較為傳統的媒介不同的是，Twitter給我們提供了一個從個人層面上研究人類關係的機會，因為你不僅能看到一個人說什麼，還能知道什麼時候說的以及說了多長時間。長期以來，比較語言學家一直在通過語言跟蹤研究群體共性。基本的單詞通常有共同的發音，比如西班牙語中的tres、法語中的trois、德語中的drei、英語中的three、印度古吉拉特語中的thran，都是「3」的意思，屬於最基本的單詞，具有類似的發音。通過研究這些現象，我們就能順著時間維度來了解人類基因和文化的變遷歷程。研究人員也曾根據Twitter用戶的語言風格對用戶進行分類。之前，曾經有語言學家研究了18.9萬名Twitter用戶發送的7500萬條推文，根據日趨鮮明的語言風格對這些用戶進行了分類。下面是我從一項早期的研究成果中摘錄出來的（見表3

—2)。 [\[12\]](#)

表3—2 根據推文類型對Twitter用戶的分類

例子	语言特征	占样本比例
nigga (黑人)、poppin (机械舞)、chillin (寒冷)	缩短词尾 (比如, er 缩短为 a, ing 缩短为 in)	14
tweetup (Twitter 会)、metrics (度量)、innovation (创新)	技术术语	12
inspiring (令人鼓舞)、webinar (在线研讨会)、affiliate (子公司)、tip (诀窍)	营销术语	11
etsy 伊蒂丝网站、adorable (萌)、hubby (老公)	手工艺常用语	5
Pelosi (佩洛西)、obamacare (奥巴马倡导的医改方案)、beck (反对奥巴马的福克斯新闻主持人 Glenn Beck)、libs (自由主义者)	党派讨论	4
Bieber (人名, 原为 Biebe, 意为“比伯”)、pleasee (原为 please)、youu (原为 you)	延长词尾 (比如, 重复词尾的最后一个字母)	2
Anipals (以 animal 和 pal 为基础的合成词, 意为“动物伙伴”)、pawesome (以 paw 和 awesome 为基础的合成词, 意为好得或坏得“令人震撼”)、furever (以 fur 和 ever 为基础的合成词, 意为“反对使用动物皮毛”)	与动物有关的双关语	1
kstew (指《暮光之城》女主角克里斯汀·斯图尔特)、robsessed (《暮光之城》演员罗伯特·帕丁森的粉丝)、twilighters (暮光迷)	与《暮光之城》系列电影有关的合成词或双关语	1

重要的是，我們要注意，這項研究對用戶進行分組時採用的標準僅僅是文字風格、聯繫對象以及信息內容。這些語言群組並不是提前決定好的。事實上，在研究人員確定為分析對象的用戶組中，上面列出的這一組是人數最多的。此外，這一組恰巧也是最健談的（即人均發送推文數量最多）和最專注的（90%的推文都是發給本組成員的）。這組成員使用的語言是特點最鮮明的，他們經常使用的100個最具代表性的單詞中，有一半都屬於「縮短詞尾」那一類。在表3—2中，你可以看到研究人員根據俚語、流行文化用詞、行話、雙關語將Twitter用戶分成了幾類。換言之，人們會因為特殊的說話方式而聚在一起，而正是這些非常重要的信息卻湮沒在了歷史中。如果你能瞭解一個人在去世前對妻子講的最後一句話，瞭解一個人在朋友面前的講話方式，那麼你就能更加深刻地認識這個人。在未來幾年裡，表3—2列出的技術專家、政治學家、營銷專家以及羅伯特·帕丁森的粉絲將相互融合，並分化重組，而我們就可以通過跟蹤研究他們在網絡上所寫的文字來了解這個融合、重組過程，這將是很有趣的。

一旦數據同語言結合在一起，那麼我們就能從時間維度去研究語言的演變，這是非常有趣的。展望未來，Twitter之類的服務是不可或缺的。目前，谷歌圖書（Google Books）這項服務正在同世界各地的圖書館開展合作，推動圖書的數字化。目前，谷歌圖書提供3 000萬本電子圖書，既有大部頭，也有小薄冊，而且這些圖書還可以檢索。這樣就防止了語料的丟失或損毀，消除了語言研究過程中的一個盲點。谷歌圖書為我們進行語言研究提供了一個龐大的數據庫，其時間跨度之長，使我們甚至能檢索到1800年的圖書，從而使我們能以更加寬闊的視野去研究語言的演變歷程，進而更加深入地審視歷史上的人類，去了解對他們重要的事物。這個數據庫催生了一個從量化角度研究文化的新領域，即「文化組學」^[13]。圖3—1中，我分析了1800—2008年每100萬個單詞提及某些食物的頻率。從這幅圖中，我們可以看出，目前被提及頻率增長最快的一種食物就是比薩，而且目前被提及頻率最高的還是比薩。所以，不妨將這幅圖命名為「比薩的現在與未來」。

每 100 万个单词提到的次数

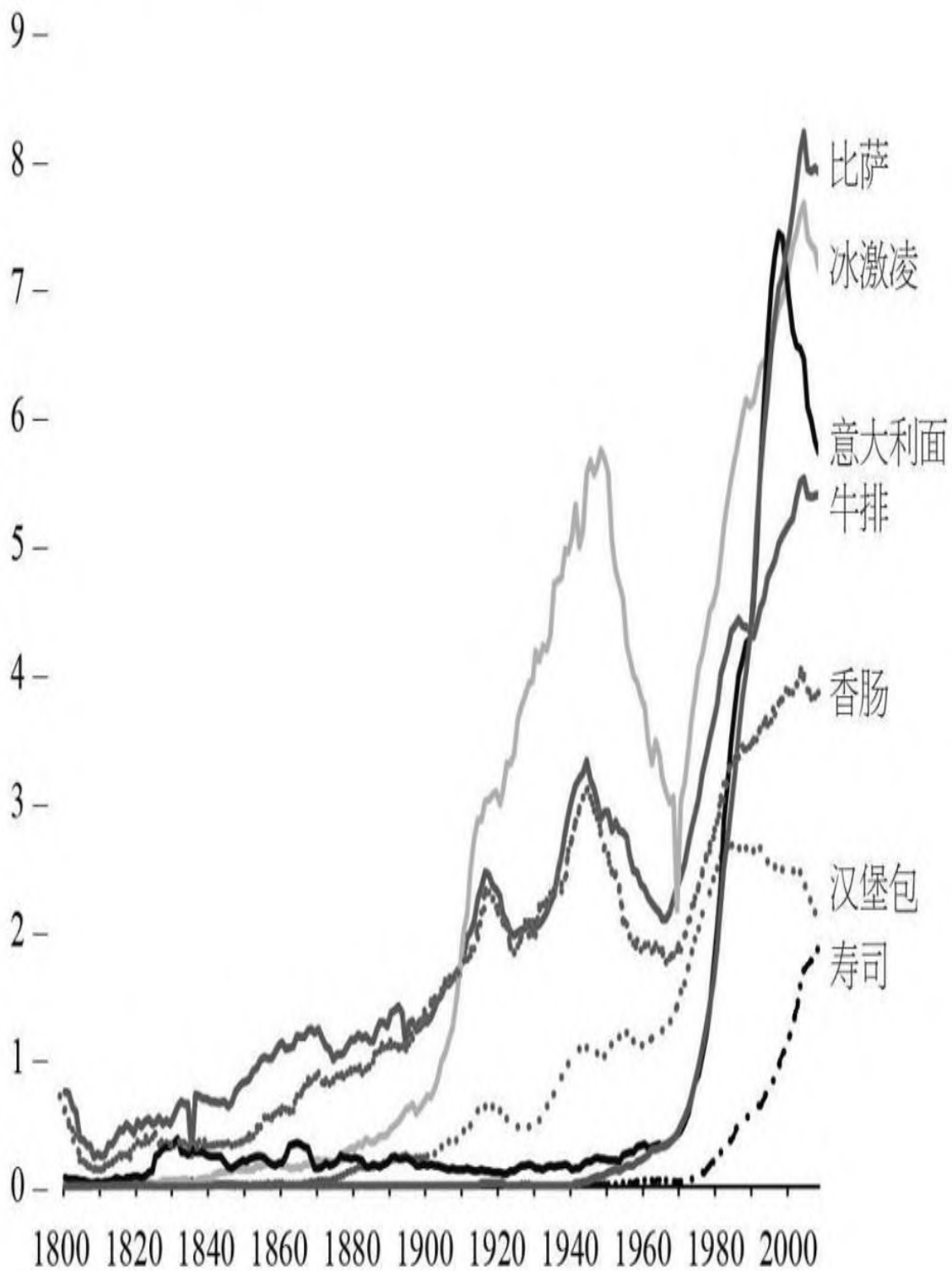


圖3—1 每100萬個單詞提及某些食物的頻率（1800—2008年）

圖3—1不僅揭示了一些與烹飪和食物有關的知識，還揭示出了一些其他方面的歷史。比如，在20世紀頭10年，冰激凌被提到的頻率之所以出現大幅攀升之勢，是因為通用電氣公司當時推出了家用冰箱。再比如，在20世紀90年代，意大利麵被提及的頻率出現了大幅下跌趨勢，這是因為當時阿特金斯健康飲食法流行起來，這種方法也被稱為「食肉減肥法」或「低碳減肥法」，要求完全不吃碳水化合物，而可以吃高蛋白的食品，即不吃任何澱粉類、高糖分的食品，而多吃肉類、魚。此外，在兩次世界大戰期間，牛排被提及的頻率也出現了大幅提升趨勢，這表明當時人類比較喜歡吃紅肉。因此，該圖說明數字技術能讓我們輕鬆地深入窺探到人類集體心理的演變歷程。^[14]詞頻研究甚至能揭示出我們如何看待抽象事物，比如時間的流逝。對於這類抽象事物，我們很難進行直接研究。問一個人「10年」對他意味著什麼，就意味著讓他描述一種顏色是什麼樣子。你原本想得到一些事實，但對方只能給你說出個大致印象。不過，如果我們能根據時間維度去研究人類書寫方式的變化，或許能獲得一定的瞭解。

數據顯示，隨著時間一年一年地流失，我們的目光往往在更大程度上盯著當前，而不是過去。比如，每100萬個單詞提到「1850年」的頻率，在第二年，也就是1851年達到頂峰，大約是每100萬個單詞提到35次。每100萬個單詞提到「1900年」的頻率最高是58次，提到「近年」的頻率最高是180次。圖3—2揭示了每100萬個單詞提到1850年、1900年、1950年和2000年的頻率的變化趨勢。

像這類基於印刷文字的分析能幫助我們在研究人類文化過程中拓寬視野，從而獲得更加廣泛深入的瞭解。Twitter讓我們看到了不同用戶群組的聚集方式，但書籍和Twitter都是一對多的交流形式，而在很多情況下，我們最重要的話語往往是藉助一對一的交流形式表達出來的，就像巴魯少校給妻子的那封信一樣。OkCupid網站的用戶每天交換約400萬條消息。當然，這些用戶交流信息是為了實現一個特殊的目的，即約會。但網站頁面並沒有給用戶提供具體的提示，也沒有設置字數或內容限制。我們可以將用戶在交友網站上交流信息的行為比作發郵件。他們最初並不認識，通過信息交流，逐漸相互瞭解，過了很久之後，開始線下見面，建立浪漫關係。但由於這些信息往往具有高度的私密性和匿名性，除了相互交流的兩個人之外，其他人無法在網站頁面上瀏覽到這些信息。為用戶發送的信息做好保密工作是一個交友網站最神聖的職責，

所以，其他研究人員無法接觸到這些私密信息。但由於我是這個網站的創始人之一，這一特殊的身份使我能夠接觸到這些數據。接下來，我就分享一下我在研究這些數據中的感悟。

每 100 万个单词提到的次数

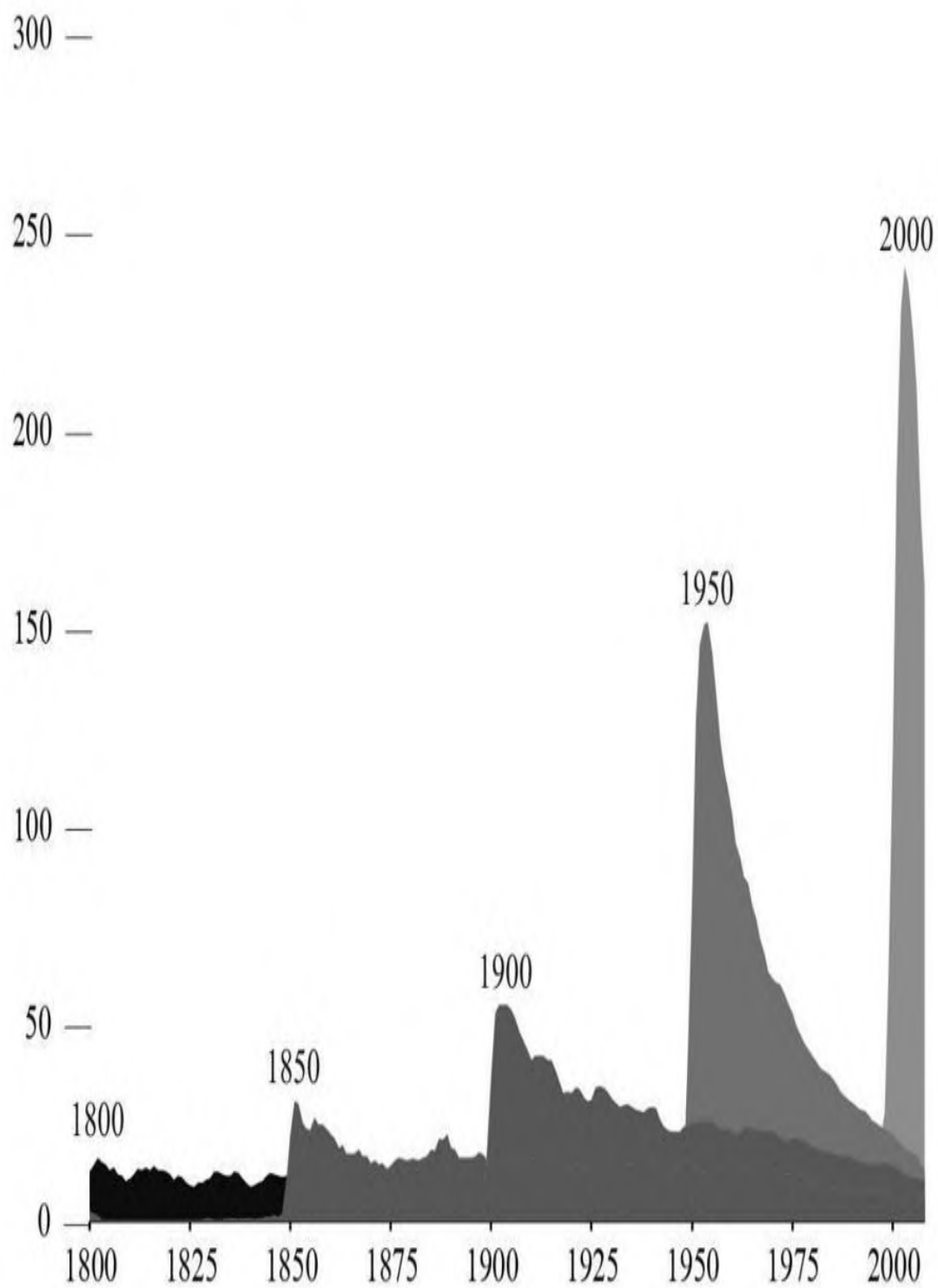


圖3—2 每100萬個單詞提到1850年、1900年、1950年和2000年的頻率的變化趨勢

首先，回顧OkCupid網站10年的發展歷程，我們就能看到技術改變了人們的交流方式。在智能手機、Twitter和Instagram產生之前，甚至在MySpace只提供文件存儲服務時，我們的網站就開始運營了，當時的一些數據一直保留到了今天。從這些年的信息來判斷，我們看到人們的書寫文化的確正在發生改變，不過這個改變是由手機推動的。蘋果公司在2008年中期正式推出iPhone手機應用程序商店。如同每一項主要的網絡服務一樣，OkCupid網站也迅速推出了一項應用程序，對人們的寫作方式產生了直接的影響。自從智能手機出現以來，用戶開始在比手掌還小的手機鍵盤上打字，信息長度縮短了2/3以上（見圖3—3）。

現在，平均來講，Twitter上的信息長度也就是100多個字符。讀者似乎已經適應了這個長度。最好的信息，也就是回覆率最高的信息，只有40~60個字符（見圖3—4）。

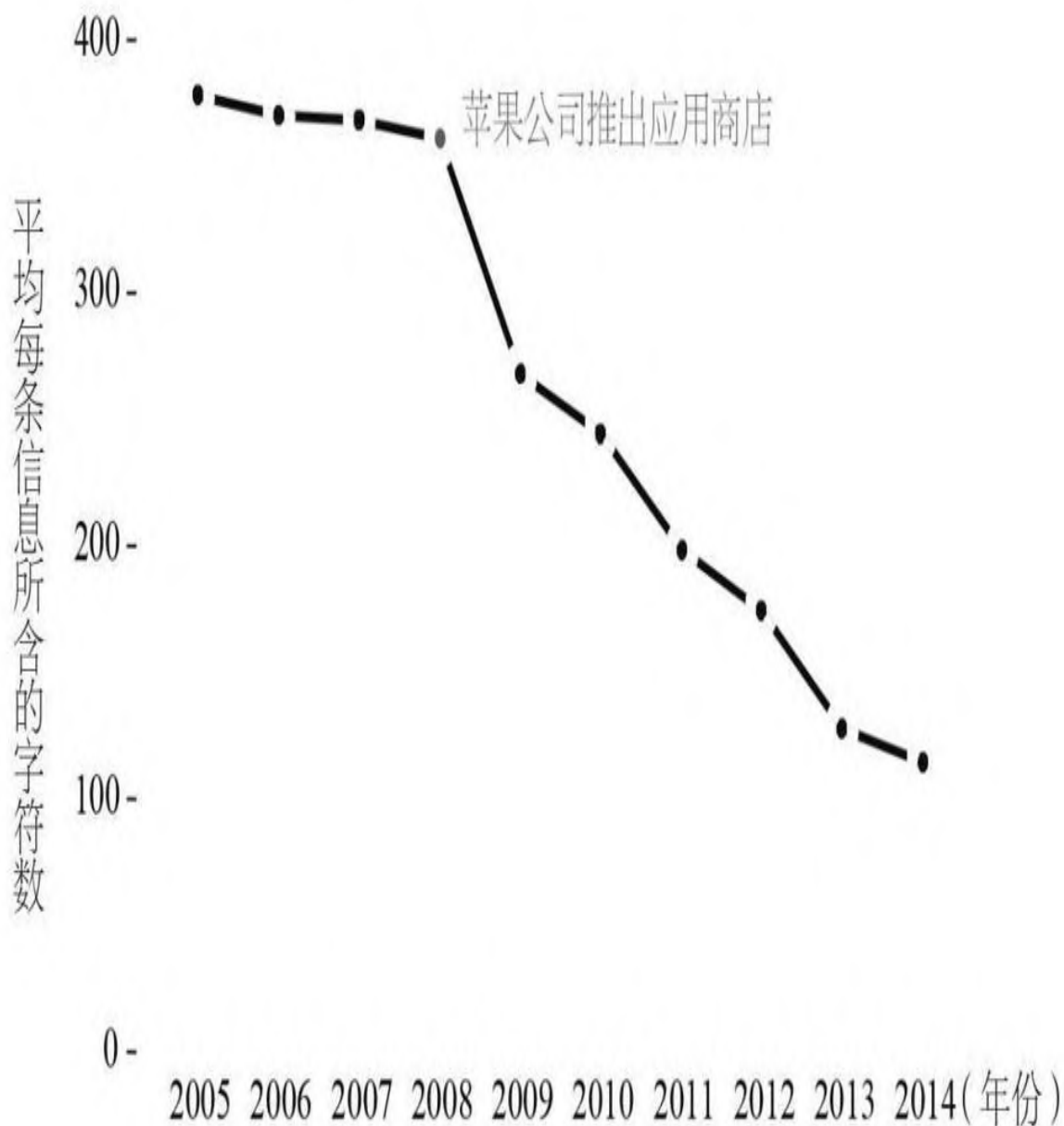


圖3—3 信息平均長度（2005—2014年）

如果審視一下特定長度的信息，然後問一下這些信息的發送者花費了多久完成撰寫，那麼我們就能看到花費多少精力、做出多少修改才能帶來更好的效果。在圖3—5中，我分析的信息長度均在150~300個字符。我分析了這些信息的發送者是花費多長時間寫出來的。正如你所看到的那樣，如果花費足夠的時間，在一定程度上會產生積極影響，有利於提高回覆率，但超過某一個點之後，回覆率會出現下降趨勢。這說

明，在寫信息時，精力並非花費得越多越好，所以，在寫信息時不要過度思考！

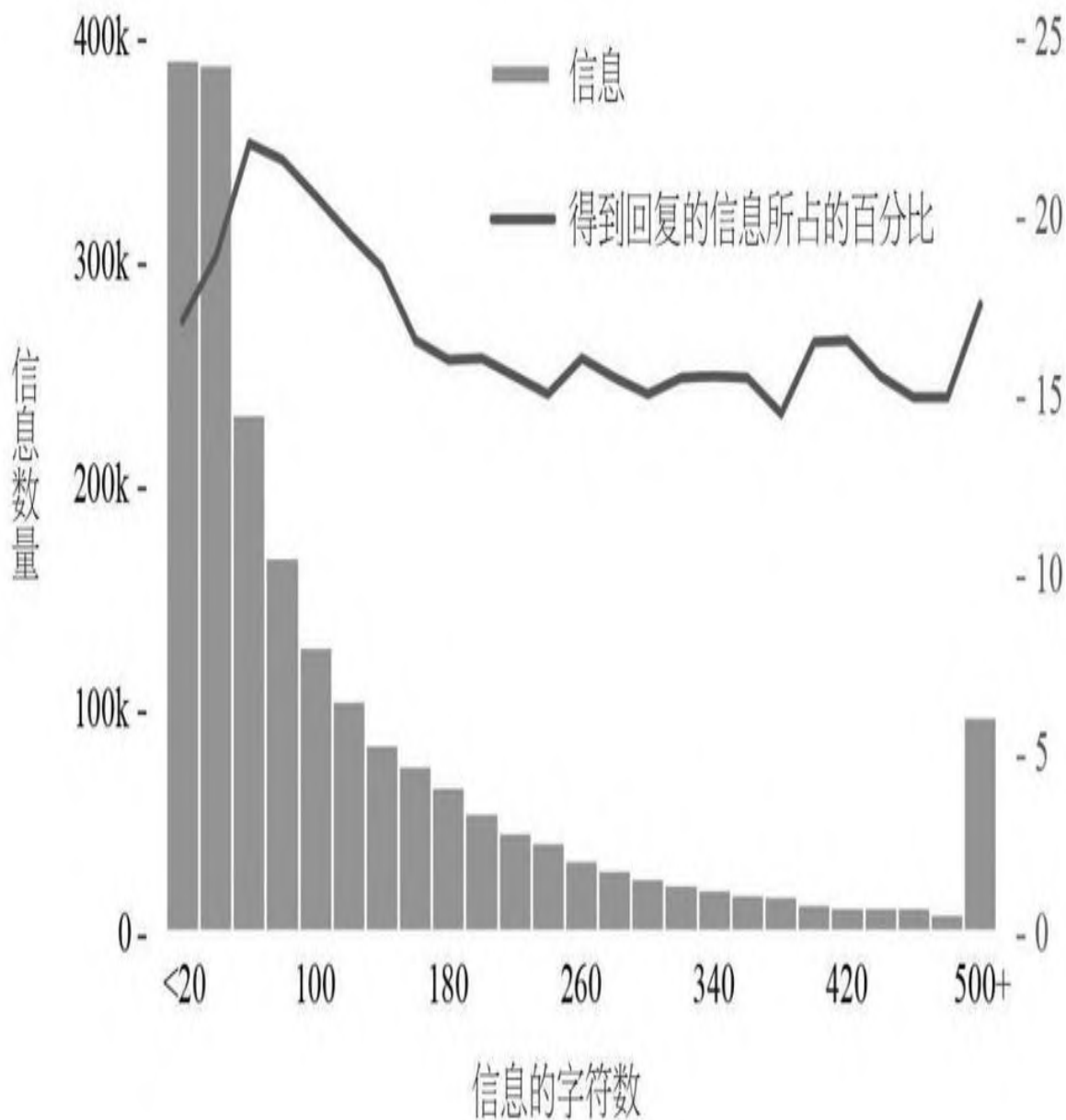


圖3—4 信息長度與回覆率

在圖3—5中，最左側的柱形表示用戶在10秒之內就完成了的信息，而恰恰是這類信息，在樣本中佔據了特別大的比例。這肯定會令很多人驚訝不已，至少我是如此。要知道這些信息的字符數都在150~300個，那麼這些用戶怎麼可能在不到10秒的時間內迅速打出如此多的字符呢？簡

短地回答，即他們並不是一個字符一個字符打出來的，有些字符是複製粘貼而來的。下面我解釋一下我是如何知道這一點的。

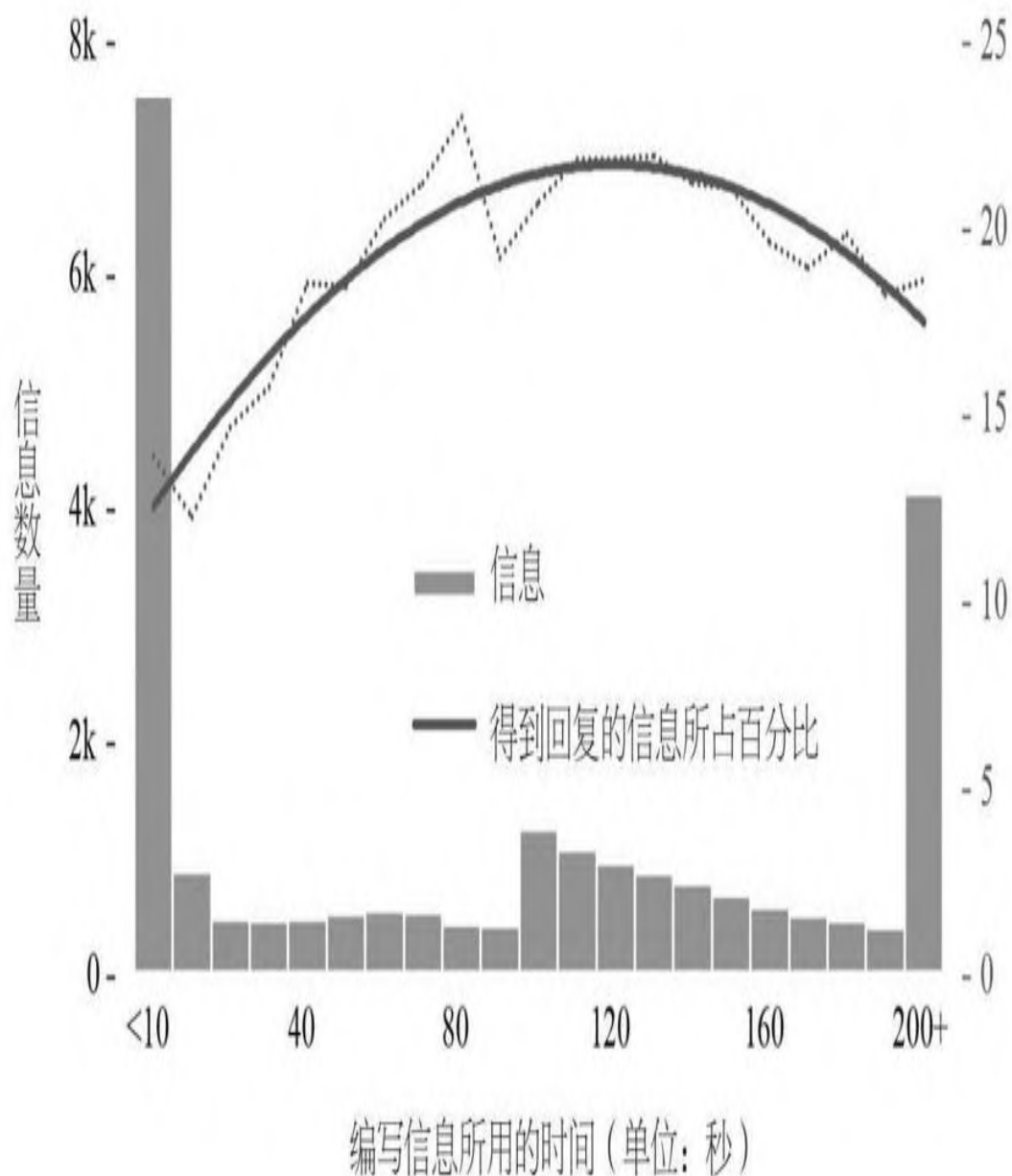


圖3—5 編寫150~300個字符的信息所用的時間與回覆率

圖3—6是我在分析了10萬條信息之後制做出來的一幅散點圖，[\[15\]](#)揭示了鍵入字符數與實際發送的信息中包含的字符數之間的關係。[\[16\]](#)研究對象的信息量太大，多達10萬條，因此，為了便於繪圖起見，我在這幅圖中標註縱軸與橫軸的數字時採用了對數法。

我在圖3—6中添加了一條對角線。如前所述，這條對角線是敲擊按鍵的次數與實際發送字符數相等的點的組合。從本質上講，發送者根據自己的所思所想打出來相應的信息之後，就直接點擊「發送」鍵，沒有刪除，也沒有修改，就將信息發送了出去。因此，我們知道信息A的發送者肯定是在匆匆忙忙的狀態下敲擊按鍵的，幾乎沒有做什麼修改就將信息發送了出去，從而使得敲擊按鍵的次數與實際發送的字符數幾乎保持了一致。在這幅圖中，有人在同陌生人第一次打招呼時，竟然花費73分41秒編寫了一則包含5 979個字符的信息，如果放到本書中，將長達4頁，結果沒有得到對方的迴應。信息B的發送者則不是這樣，因為他在第一次跟陌生人打招呼時，按鍵387次，結果實際發送的信息中包含的字符數還不到10個，付出巨大勞動之後還能讓自己的信息保持如此簡潔。

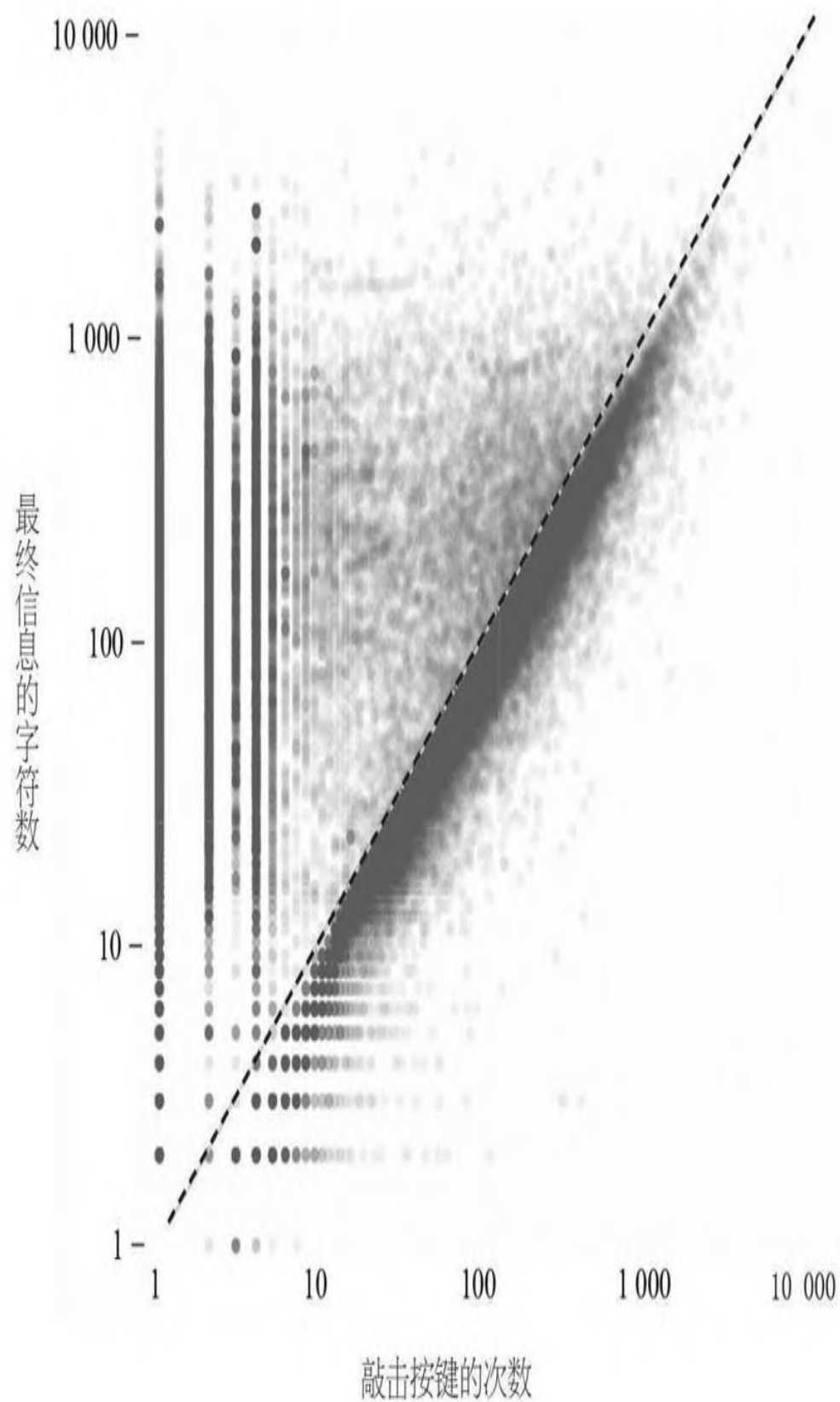


圖3—6 透明處理圖

但信息A和信息B都是極端的例子。這幅散點圖要傳遞的普遍意義在於：一條信息越趨向於對角線，修改和編輯得越少；越趨向於右下方，對信息進行的編輯越多；越趨向於左上方，就說明覆制粘貼得越多，因為敲擊次數太少，而實際發送的信息包含的字符數太多，如果沒有複製和粘貼，最終發送的信息幾乎不可能包含這麼多字符。

我們可以對圖中的每個點進行透明處理，將其透明度變成90%，這樣你就能更加清楚地看到這些點的密度（見圖3—7）。這些數據似乎經過了X射線的照射一樣，使我們透過表面看到其內部的骨架。

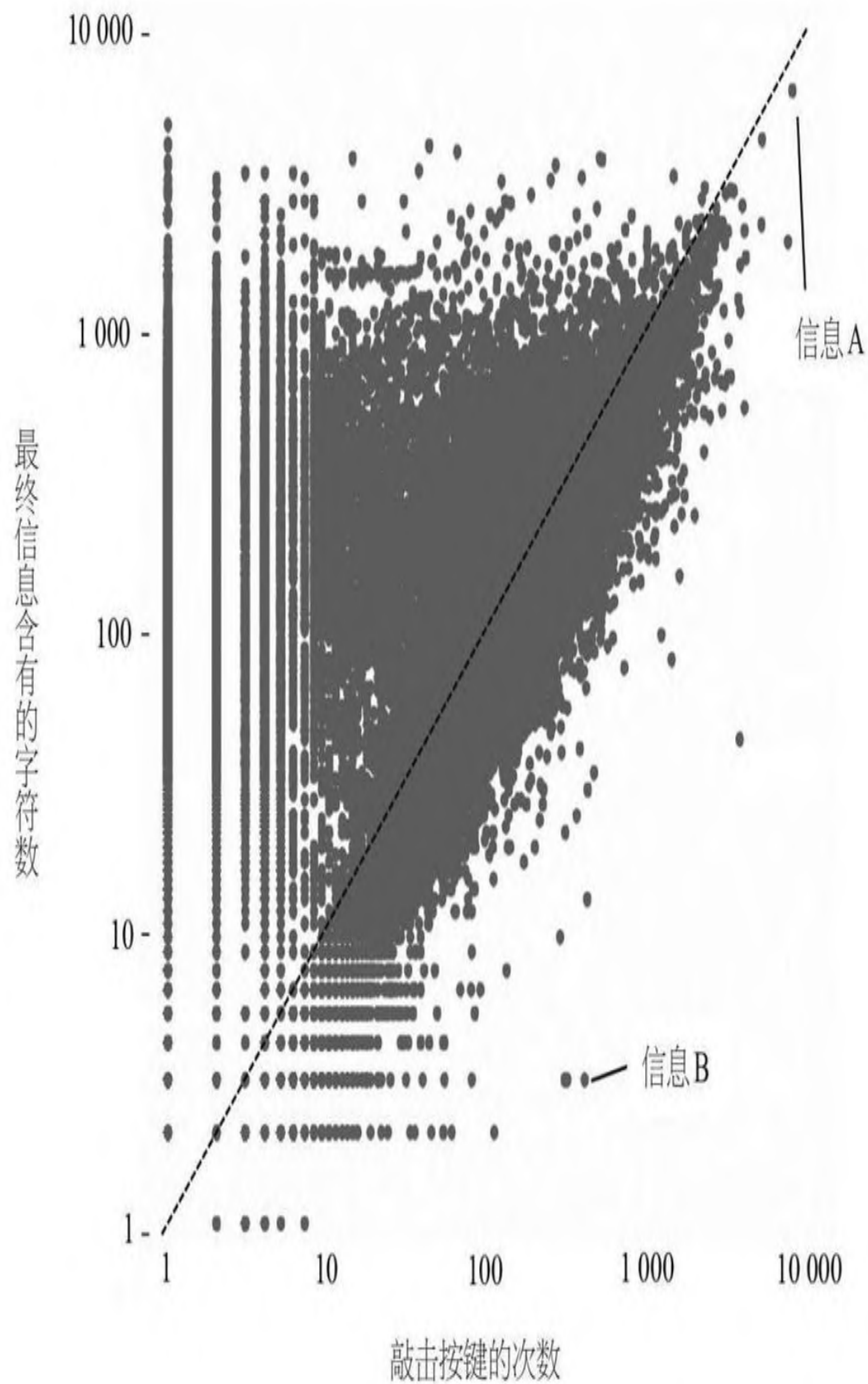


圖3—7 鍵入字符與實際發送的字符

對角線右下方點的密度是比較高的，顯得非常密集，令人驚訝。這些點代表信息真正是敲擊按鍵打出來的。當然，在這條對角線左上方的點代表的信息是粘貼而來的。因此，這條對角線就把這兩類信息分隔開了，就像把兩個對立的派別分隔開一樣。但這條對角線最左下側的信息非常稀少。對於人們究竟願意付出多少精力來編寫出一條信息，似乎存在一個天然的界限。做一下數學運算，你就會發現，最終信息中的字符數與實際敲擊按鍵的次數之比為1:3。

位於對角線左側的那些信息發送者認為沒必要花費太多精力，他們往往是對一個信息模板稍作修改後就發送了出去。由於縱軸與橫軸是按照對數法標註的，所以，這幅圖可能讓你感到很困惑，因為對角線左側信息中的大部分內容都是模板上的。在對角線左側，往左看，就會發現一道道密度較高的線條，如同車轍一樣。事實上，這些信息的發送者只是隨便按幾下鍵，對固有的模板稍作修改，就成了最終的信息。在我的樣本中，這類信息的數量還不少，20%的信息敲擊按鍵的次數都不超過5次。這些人鎖定了一些自己喜歡的信息模板，或者找到了一些自己認為可能用得上的模板，索性就拿來用了。但這類信息並不是我們平時收到的垃圾郵件，如果OkCupid網站上存在假冒的賬戶，或者專門發送垃圾信息的木馬程序，我們的管理員就會迅速將其封掉。這些信息都是我們的用戶為了聯繫陌生人而發送的真實信息。你肯定已經猜到了，他們的確是普通的懶人。他們與異性打招呼時，只是簡單地說一句：「嗨，你好可愛啊。」「想聊聊嗎？」「經常來這兒嗎？」但有些重複使用的信息具有一些特點，很難相信這些信息會同時適用於多人。下面就是一個例子，我原封不動地附在這裡：

我有抽菸的習慣，這是5月份揹包旅行時養成的。以前睡醒後總想喝點什麼，但現在醒來後竟然想抽支菸。有時候，我很希望能在《廣告狂人》那種辦公室裡工作。你看過紐約現代藝術博物館舉辦的勒·柯布西耶作品展嗎？聽起來非常有趣。就在上週，我還到蒙特利爾看了建築師弗蘭克·蓋裡的作品展，領略了他如何用計算機模型在俄亥俄州設計了一座非常漂亮的房子。[\[17\]](#)

這就是那條完整的信息。這個人試圖聯繫既抽菸又熱愛藝術的女性。對於他這種坦率的、非矯揉造作的態度，我也是非常喜歡的。網站的數據顯示，此人將這條信息原封不動地發送給了他喜歡的42名女性。

根據整個網站的數據來分析，複製粘貼的模板信息不如原創信息有

效，前者的效果要比後者低25%。但如果從付出與回報來看，前者往往是贏家，也就是說，從每付出一個單位的時間獲得的回覆量來判斷，前者比後者的效率高很多倍。我曾經把有些傢伙複製粘貼的事告訴過別人，結果對方總是說「這太拙劣了」。我說模板的效果比原創信息低25%，他們很懷疑，因為他們相信如果信息是複製粘貼過來的，那麼幾乎每一個人都能識破。但我在前面列出來的那條信息卻很高明，雖然原封不動地發給了多人，卻不大可能露餡，而且耗費的時間比原創信息少得多。42名女性收到這條信息，有5人進行了回覆。讓我告訴你一件事。我桌子上的每一件物品、我身上的每一件衣服以及我家裡幾乎所有物品，都是由工廠按照固有的模板生產出來的，誰知道這個工廠生產了多少件。每次吃午飯的時候，我都要在一堆人裡擠來擠去，從擺成了一堵牆似的三明治裡拿出來一個。由此來看，模板和其他模式化的事物是有用的。我們在前面提到的那個喜歡抽菸的揹包客在發送求交往的信息時，雖然採用了複製粘貼的辦法，但他的所作所為與人們在其他領域的做法具有異曲同工之妙：大家都在利用技術來為自己服務。在這種情況下，他做出的創新之舉無非是少敲幾下按鍵、節省一些時間而已。

正如我們所看到的那樣，很多智能手機以及Twitter等網站在提供服務時，往往會提出自己的各種條條框框，讓用戶去適應，但寫作就像人類生命一樣，是永恆存在的。寫作方式可能會改變，人們也可能用一些奇怪的複製手段，也可能出現其他一些出乎意料的變化，甚至如同一切有生命的事物一樣，偶爾也會出現退化，但我們必須認識到這樣一個事實，即目前，寫作無異於正在經歷一次寒武紀生命大爆發時期，而不是大規模滅絕時期。現在，雖然人們可以從剪貼板上直接複製文字，但語言的多樣性特徵比以往任何時期都顯著。多樣性有利於一種文字和藝術的保護，而不會對其發展構成威脅。從文學著作中那些陽春白雪般的語言到社交網絡上那些簡單的甚至存在錯別字的狀態更新，所有這些寫作都貫穿著一個共同的目標。無論是從朋友到朋友，還是從陌生人到陌生人，無論是從情人到情人，還是從作者到讀者，彼此之間賴以建立關係的紐帶都是語言。當一個人無聊、激動、憤怒、欣喜、戀愛、好奇、想家或擔心未來時，他都會將自己的想法付諸筆端。

[1] 因為這個現象非常有趣，而且令人感到很驚訝，涉及的資料比較複雜，所以我參考了許多數據來源。下面是直接引用的英文文獻：「Dying to Go Home,」by Jackie Rosenhek, Doctor's Review, December 2008, doctorsreview.com/history/dying-to-go-home/. 「Beware Social Nostalgia,」by Stephanie Coontz, New York Times, May 19, 2013, nytimes.com/2013/05/19/opinion/sunday/coontz-beware-social-nostalgia.html. 「When Nostalgia Was a Disease,」by Julie Beck, The Atlantic, August

2013,theatlantic.com/health/archive/2013/08/when-nostalgia-was-a-disease/278648/.The
「Nostalgia」entry on qi.com: qi.com/infocloud/nostalgia.

[2] Facebook在2013年第四季度的營業收入導致人們質疑它的表現。詳情請參考喬安娜·斯特恩（Joanna Stern）於2013年10月31日在美國廣播公司新聞網發表的文章，名為《青少年離開Facebook後去了哪兒》（Teens Are Leaving Facebook and This Is Where They Are Going），鏈接：abcnews.go.com/story?id=20739310。

[3] 我在美國證券交易委員會保存的Facebook公司季度報告裡瞭解到了孩子們喜歡什麼。我真是太瞭解現在的孩子了。

[4] 關於他這封信的基本信息，可以參考下面這個鏈接：pbs.org/civilwar/war/ballou_letter.html。雖然他這封信在生前一直沒有寄出，但在其去世後，這封信和其他遺物一併交給了他的家人。

[5] 我的估算依據如下：至少根據谷歌的數據，目前人類所寫的書的數量為129 864 880種。這個數字精準到有些可笑的地步，但鑑於谷歌已經收錄了其中3 000萬種，並且對其進行編目是谷歌的專業特長，因此，谷歌的評估大致是可信的。具體情況請參考鏈接：mashable.com/2010/08/05/number-of-books-in-the-world/。根據亞馬遜的數據，一本小說字數的中位數是6.4萬個單詞。因為這個中位數與平均數之間的差距不會很大，所以，我就假設平均數也是這個數字。在我看來，小說和其他類型書籍的字數差別應該也不會太大，具體請參考蓋布·阿布什（Gabe Habash）於2012年3月6日發表的文章，鏈接為：blogs.publishersweekly.com/blogs/PWxyz/2012/03/06/the-average-book-has-64500-words。根據這兩個數字來計算，紙質書籍的總字數就是8 311 352 320 000字。Twitter在2013年8月報告說，用戶一天總共發出5億條推文，請參考：blog.twitter.com/2013/new-tweets-per-second-record-and-how。我估算每條推文大約擁有20字，也就是每天100億個字。這樣來算，Twitter只要831天（2.3年）的時間，總字數就能超過所有紙質書籍的總字數。顯然，這只是我的估算，而且是很保守的估算，因為現在越來越多的人每天會發好幾條推文，很有可能導致Twitter的總字數提前超越紙質書籍的字數。

[6] 費因斯這句話得到了廣泛引用。詳情請參考盧西·瓊斯（Lucy Jones）於2012年10月27日在《每日電訊報》上發表的Ralph Fiennes Blames Twitter for ‘Eroding’ Language一文，鏈接如下：telegraph.co.uk/technology/twitter/8853427/Ralph-Fiennes-blames-Twitter-for-eroding-language.Html。

[7] 我在此處和其他地方對Twitter所做的分析中，都是基於我的研究團隊隨機蒐集的120個賬戶的用語，形成了一個具有代表性的語料庫。

[8] 關於牛津英語語料庫及其最常用的單詞，可以通過下面這個鏈接瞭解到更多信息：en.wikipedia.org/wiki/Most_common_words_in_English。牛津英語語料庫只列出了動詞原形，而沒有列出動詞變位的情況。比如，只列出了have，而沒有列出had、having和has。此類情況有很多。我在研究Twitter用語時決定不這麼做。雖然我的決定導致直接比較列表變得比較困難，但我比較傾向於儘量展現出這些數據的原始面貌。

[9] 利伯曼教授的博客名稱為Language Log，鏈接為：languagelog.ldc.upenn.edu/nll/。這個博客中有很多有趣的文本分析案例，尤其是他對前面提到的拉爾夫·費因斯那句名言的分析。請參考該博客於2011年10月28日的博文，名為「Up inUR Internets, Shortening All the Words」，鏈接為：languagelog.ldc.upenn.edu/nll/?p=3532，尤其是有關費因斯那句名言的討論值得關注。

[10] 在分析詞語長度之前，我和利伯曼已經排除了所有的網址以及「@」和「#」這兩個特殊符號，因此不必擔心這些成分增加單詞長度。

[11] 這一段關於Twitter的其他文本分析引自Yuheng Hu、Kartik Talamadupula與Sub-barao Kambhampati在美國人工智能協會於2013年7月8~11日召開的馬薩諸塞州坎布里奇市召開的第七屆博客與社交媒體年度國際會議上提交的論文，論文題為「Dude, srsly?: The Surprisingly Formal Nature of Twitter's Language」，論文獲取鏈接如下：

aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6139。

[12] 這個表格以及後續關於Twitter詞彙的討論，引自John Bryden、Sebastian Funk與Vincent AA Jansen在EPJ Data Science（2013年第3期）上發表的「Word Usage Mirrors Community Structure in the Online Social Network Twitter」一文。我還引用了他們的附加材料，這份附加材料包含了一些沒有列入論文討論範圍的原始詞群列表。完整的論文以及附加資料的獲取鏈接為：epjdatascience.com/content/2/1/3。

[13] 這種通過挖掘谷歌圖書探尋文化趨勢的方法最早出自讓—巴蒂斯特·米歇爾（Jean-Baptiste Michel）等人於2011年在《科學》雜誌上發表的一篇文章，「Quantitative Analysis of Culture Using Millions of Digitized Books」，載於Science 331, no.6014（2011）：176—82, doi:10.1126/science.1199644。關於各個時代的食物的名詞，我就引用了這篇文章的研究成果。這篇文章提到人類的記憶力有「半衰期」，我對此不敢苟同。但作者所說的「我們遺忘過去的能力逐年增強」，顯然在大方向上是正確的。這篇文章有趣的地方不僅侷限於我提到的這兩張表，所以值得完整地讀一遍。

[14] 谷歌圖書的數據庫考慮到了當前書籍出版規模遠遠大於從前（比如19世紀），因此，每年在抽取樣本時，選取的圖書數量是固定的。正是由於這個原因，雖然我們這裡的兩張圖表都揭示出有關詞語的詞頻越來越高，但這的確是因為人們對這些詞語的興趣越來越大，而不是因為現在出版的書籍增加了。有的詞語的詞頻非但沒有隨著時間的推移而提高，反而有所降低，比如「God」（上帝）。這個詞的詞頻在過去10年內持續下降，在當前的美國書面文本里，詞頻大概只有19世紀伊始的1/3。culturomics（文化組學）這個詞是由讓—巴蒂斯特·米歇爾（Jean-Baptiste Michel）和埃雷茲·利伯曼·艾登（Erez Lieberman Aiden）這兩位學者在《基於數百萬電子圖書的文化量化分析》（Quantitative Analysis of Culture Using Millions of Digitized Books）一書中率先提出來的。這裡的圖表和數據都是參照他們的研究成果編制的。

[15] 我們在分析的過程中，任何人都不會去閱讀隱私的信息。OkCupid一直有檢測垃圾信息的軟件，平時會對用戶進行抽樣檢測，自動記錄敲擊鍵盤的次數和打字時間。因為我沒有讀到任何隱私內容，我這裡所說的只有三個字母的私信內容是hey，其實只是我猜測的一種可能性，而不具有肯定性。在OkCupid網站上，三個字母的信息裡，大約80%都是hey。緊隨其後的是sup，然後就是wow。鑑於hey的受歡迎程度遠遠超過其他三個字母的單詞，而且我只是在開玩笑，無論採取其他任何一個單詞，其實都沒有實質意義上的區別，因此，我最後決定選擇hey。

[16] 這裡的鍵入字符數是通過我為本章設計的程序而獲得的。

[17] 這篇逐字逐句引用的個人信息是我在其他網站上看到的，引起了我的關注。我是得到作者的同意之後才在這裡引用的。

第四章 社交圖譜

通過約會網站收集的數據具有一大缺點：對於網友線下約會的情況，你幾乎無法獲得任何有用的信息。一旦約會雙方見了面，就不需要信息、評級或其他類似的東西了。這簡直是對數據集和這份工作本身的雙重諷刺——你什麼都沒做錯，而用戶卻離你而去。好在用戶成雙成對了！

他們會去哪兒呢？當然是去現實世界，去酒吧，去陽光下，也有可能發生性關係。簡單地說，他們離開被比特和像素輕鬆量化的世界，進入了各自的生活。想想一對年輕人的交往是如何進展的吧。兩人第一次見面，聊天，喝酒，瞭解彼此。下一步，如果還有下一步的話，便是各自的公寓了：門上陌生的房號，銅把手，你家的是鋼的。別人床單奇怪但好聞的味道。浴室裡的洗髮水，用過的，但是對你來說是陌生的。羅甘莓：好啊，為什麼不呢？等到下次到你家的時候，她打開冰箱，發現只有……芥末醬。抱歉。到別人的家裡之後，開始聊形形色色的事和人，開始對一些小玩意兒感嘆不已，緊接著，你們開始聊起了一項游泳邀請賽，事實上你可能正是通過她才得知這個賽事的。

你們見各自的朋友，把對方介紹給自己最好的朋友，就像你們已經認識很久了一樣。足夠的酒精，合適的人群，他們也會變成你的朋友。對方的點頭之交和同事，有心無心地，漸漸也認識了。最後，也許，如果你們之間的關係真的要修成正果了，雙方父母便參與進來了。你們當著大家的面，講述人生故事更精彩的一面，有一部分是你們可以共同描述的，因為你們是如此熟悉對方。離開桌子一會兒，父母知道的事情會比你離開的時候更多。回到你的座位，「媽媽告訴我……」，這是你講述自己最喜愛的故事的完美開場白。兩個人的生活不斷融合。然後，經常出現的一種情況是，關係戛然而止，一切又要從頭開始，與另一個人重新開始。

以上便是兩個人在第一次被對方吸引後成為戀人的種種方式。我認為電腦絕不可能將他們最終在一起的過程全部捕捉下來，但是我們能夠預測他們在一起之後的生活。情侶的生活模式，也就是對他們所謂的「社交圖譜」進行的記錄，已經十分完備。

我在Facebook上有384個朋友。在圖4—1中，每一個圓圈代表一個朋友，中間那個點代表我自己。^[1]我右側那個黑色的點代表我的妻子萊西瑪。連接各點的灰色線條代表著人與人之間的聯繫。

圖4—1很好地展現了我的朋友圈，但它不是手工製作的，而是我一個非常能幹的研究助手詹姆斯·多戴爾編寫了一款特殊的軟件，用這款軟件製作出來的。這款軟件根據共同好友的數量對我的好友進行劃分，哪兩個人的共同好友多，那麼代表他們的那兩個圓圈就靠得近。我們可以將這些小圓圈想象為「鐵粉」，這些「鐵粉」被友誼的力量磁化之後，落到桌面上固定了下來。我只通過Facebook接受朋友的邀約，除此之外，用得並不是很多，但你仍然可以從圖上發現我生活中的聯繫人網絡。在圖4—1中，我的姻親之間聯繫得非常緊密，一條條灰色的線條相互重疊，幾乎達到了這款軟件允許的最大密度，這個聚集區被標為A區。我的中學同學的聚集區被標為B區。我的同事的聚集區被標為C區。和我一起遊戲、運動的朋友的聚集區被標為D區。從這幅圖中，你甚至能讀出來我曾經的和未來的音樂生涯。我在一個樂隊裡待過幾年，跟著樂隊到處巡迴演出。^[2]該圖左側那些相對孤立的圓圈主要是我在演出之路上遇到的人。他們彼此之間賴以維持關係的紐帶是音樂，我的算法無法體現他們之間的聯繫。

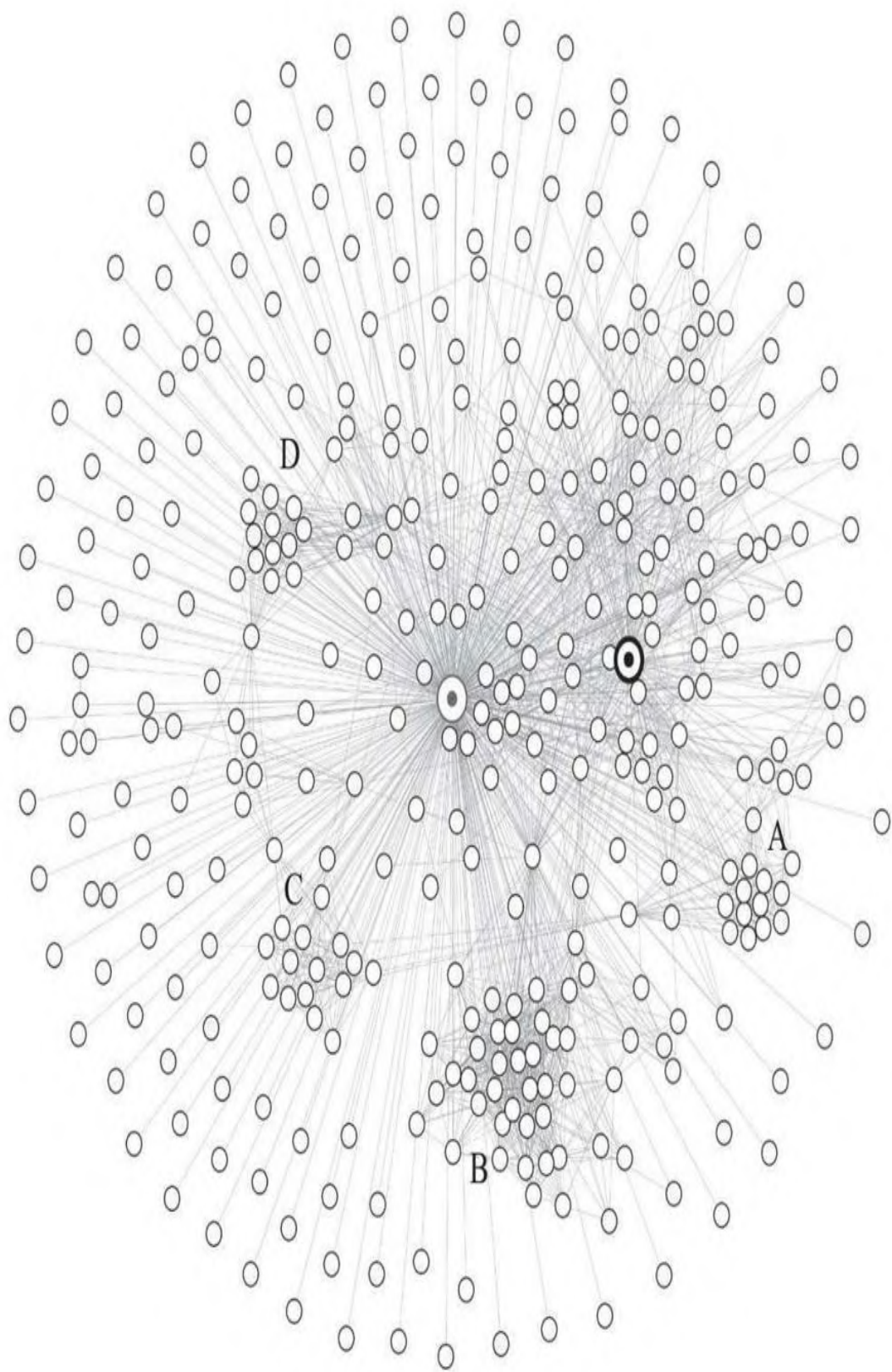


圖4—1 我的社交圖譜

在圖4—1這幅社交圖譜的基礎上，我進行了拓展，將萊西瑪社交圖譜與我的社交圖譜合併在了一起，顯示出了我們作為一對夫妻的社交網絡範圍（見圖4—2）。深紅色的點表示我和妻子共同的朋友。

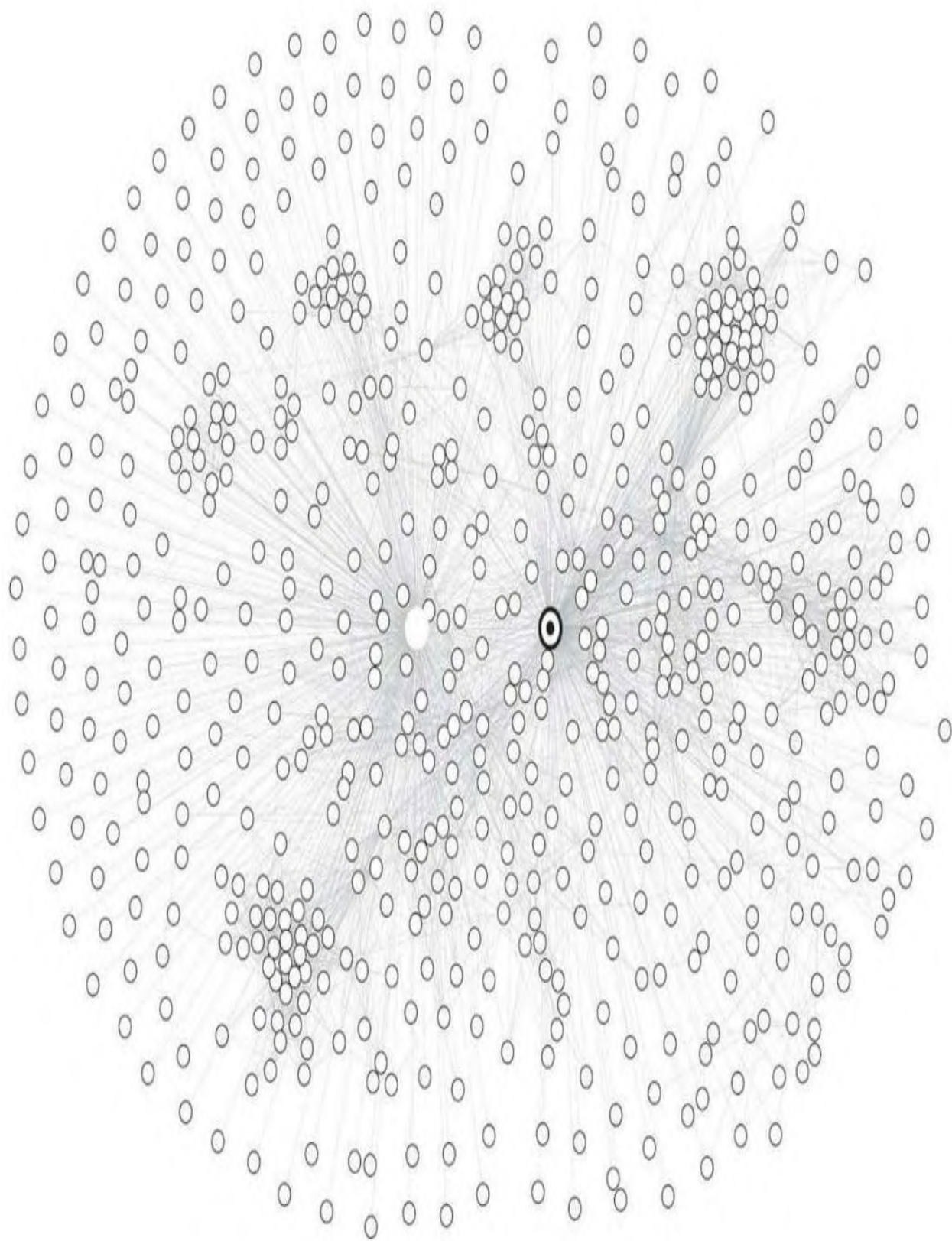


圖4—2 我和妻子的社交圖譜

這幅圖以枯燥而抽象的方式描繪了我和妻子的朋友圈。從該圖中，你可以看出我們兩人生活的交集。僅僅從這幅圖來判斷，你就可以看出我和萊西瑪離婚的可能性會遠遠低於其他夫妻。

像上面這種以點和線為基礎的分析方式就是網絡分析。網絡分析並非一門新興的科學，它已經存在了將近300年的時間。在網絡分析逐漸發展的過程中，人們可以用來分析的數據的數量也從涓涓細流變成了滔滔洪流。人類歷史上有據可查的第一個網絡分析問題就是「哥尼斯堡七橋問題」。18世紀初，在普魯士哥尼斯堡鎮（今俄羅斯加里寧格勒）流傳著這樣一個問題：有七座橋將普雷格爾河中兩個島及島與河岸連接起來（如圖4—3所示），是否可能從這4塊陸地中任一塊出發，恰好通過每座橋一次，再回到起點呢？萊昂哈德·歐拉（Leonhard Euler）在1735年訪問普魯士期間著手研究這一問題。^[3]他在《哥尼斯堡的七座橋》一文中闡述了他的論證過程。他把現實問題簡化為平面上的點與線組合，每一座橋視為一條線，橋所連接的陸地視為點。他將整個鎮子視為一個網絡，一個新學科就此應運而生。

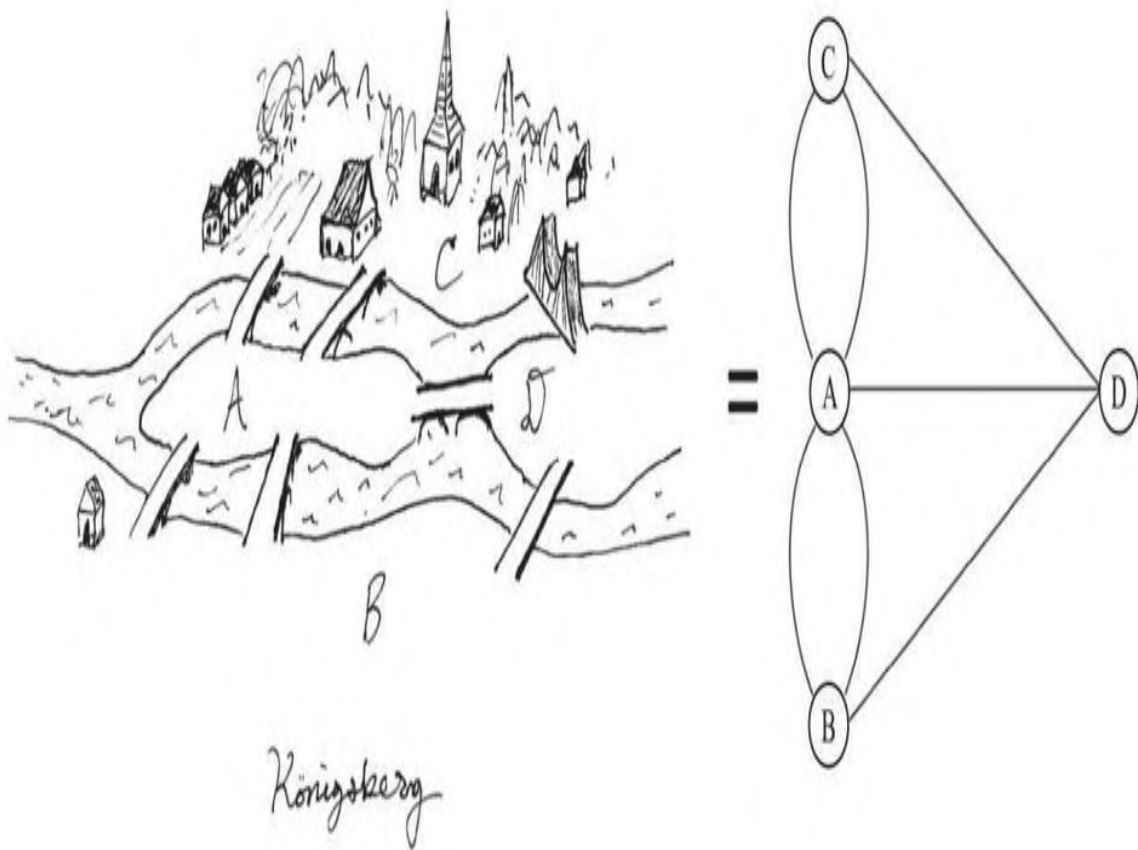


圖4—3 哥尼斯堡七橋問題

他的論證過程是這樣的：由於每座橋只能走一次，所以如果從某點出發後最後再回到這個點，那麼這個點必須有偶數條連線，也就是有來路必有去路，這樣的點稱為偶頂點；相對地，連有奇數條線的點稱為奇頂點。因此，要解答這個問題，就簡化成了在整個網絡中尋找沿途的每一個點是否有偶數條連線（即偶數座橋）。由於哥尼斯堡七橋問題中存在4個奇頂點，所以無法實現之前設想的走法，從而解答了這個問題。這個以生活小事為論證對象的過程開創了數學的一個新的分支——圖論。這種科學直到今天才發揚光大，但在我看來，這恰恰是人文精神的體現。^[4]歐拉這種節點和邊的概念原本只是為了解決日常生活中的走路問題，但後來卻幫助我們理解了很多其他問題。^[5]今天，圖論得到了廣泛應用，可以用來研究疾病與其傳播路徑、卡車與其行駛線路以及基因與其黏合酶，當然，也被用來研究個人及其人際關係。在過去幾十年間，網絡理論的運用呈現出了爆炸式發展的態勢，因為網絡本身已經大

大拓展了。

40年前，斯坦利·米爾格拉姆（Stanley Milgram）^[6]向內布拉斯加州奧馬哈市的100個人郵寄包裹（附有任務指示和郵資已付的信封），進行「六度分隔理論」實驗，希望也許幾十個有冒險精神的人能夠參與進來。通過這種別出心裁的方式，他提出了一個著名的六度分隔理論，但是並沒有證實該理論。2011年，Facebook史無前例且聲勢浩大的規模讓我們意識到米爾格拉姆是對的：當時7.21億個賬戶中，99.6%只需6步或更少便能相互關聯上。^[7]

如今，由於數據處理技術日趨成熟，網絡理論展示了人們應該如何找工作，如何從無稽之談裡提取有用信息，甚至如何製作更好的電影。在建立皮克斯總部的時候，他們做了一個大膽的決定：把大樓裡的唯一廁所設在中庭，促使部門間的人私底下聊聊天，因為他們知道創新來自偶然間的思想碰撞。^[8]這便是將「弱聯繫的力量」^[9]應用到實際，這個假想於20世紀70年代提出，因新興的、活躍的網絡數據而大放光彩：它告訴我們，正是那些我們不太熟悉的人，尤其是新認識的面孔，能幫助我們傳播點子。^[10]

網絡理論中另一個長久以來形成的觀點是「嵌入度」^[11]。嵌入度的一種表現形式是兩個人社交圖譜的相交程度。簡單來說，萊西瑪和我的嵌入度就是我們兩人圖譜中相交部分與整個圖譜的比例。利用各種資源（電子郵件、即時通信、電話）進行的研究表明，兩個人的共同朋友越多，他們的關係就越牢固；關聯越多，意味著他們待在一起的時間越多，關係越穩定。但是與電話記錄或電子郵件等不同的是，線上社交網絡為圖譜的邊緣地帶和節點提供了豐富的數據（比如約會地點如何亙古不變地見證著追求儀式，年齡增加和美貌也是研究的變量），當然，Facebook是這類網絡中信息最豐富的。人們正在逐漸感受到這種效應。

社交圖譜分析最開始關注的是「誰認識誰」，現在大抵仍然如此。Facebook的數據容量之廣——你可以不費吹灰之力便認識六度以外的人——正在顛覆這種現象。對於人際關係，尤其是情侶關係，近來這些數據催生了一種新的有效的方式，來衡量兩人之間的紐帶有多牢固。結果證明兩個人的生活應該不僅密不可分，而且以特定的方式緊緊相連。另外，網絡分析評估罕見地將「誰不認識誰」作為重要的考量。

美國康奈爾大學的計算機科學家喬恩·克萊因伯格（Jon Kleinberg）與Facebook的高級工程師拉爾斯·貝克斯特倫（Lars Backstrom）在2013

年發表了他倆合著的一篇研究論文。論文指出，在研究了超過130萬對夫妻之後，他們發現夫妻二人之間擁有的共同好友越多，這對戀人就越有可能分手。下面，我們用兩幅圖來解釋一下他們的觀點（見圖4—4）。A和B代表一對夫妻。左圖看似是比較理想的情況，夫妻二人的朋友圈交集非常多，嵌入度非常高，但婚姻關係較為牢固的卻是右圖所示的夫妻，因為一方是另一方在社交世界中的一座橋樑，這樣的夫妻關係更為牢固。

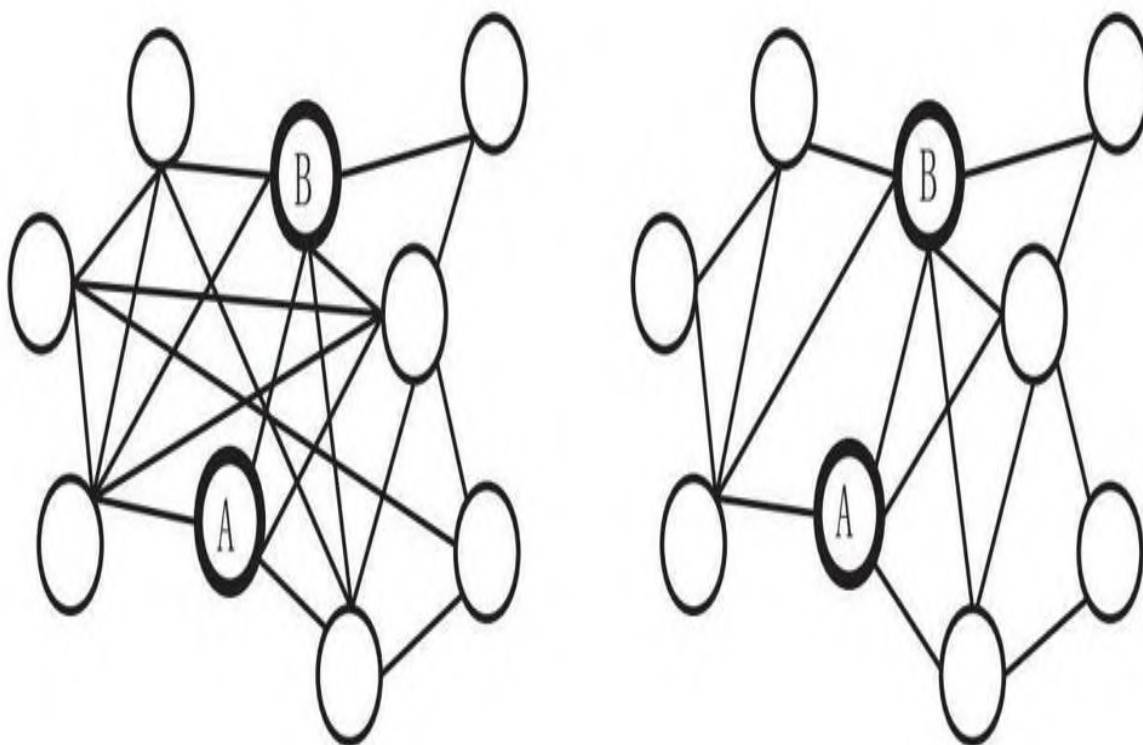


圖4—4 兩種夫妻朋友圈交集

他們的理論聽起來似乎有些奇怪，為什麼要讓你和愛人的朋友圈適度分散呢？如同一些優秀的觀點一樣，我們在現實生活中也能本能地感覺出這種觀點的正確性。以我和我妻子萊西瑪為例。萊西瑪的表弟希爾在她生活中的嵌入度就非常高。他們兩個是一起長大的，他，像她一樣，幾乎與大家族的每個成員都有聯繫，其中許多人我都不認識。他們從小就認識，而萊西瑪和我才結婚7年。如果以希爾和萊西瑪為中心繪製一個社交圖譜，則非常類似於我在圖4—4中給出的左圖。然而，希爾不認識萊西瑪的同事，不認識萊西瑪的舞蹈團成員，也不認識萊西瑪的大學同學。但這些人我都認識，更重要的是，這三個群體的交集中只有

我一個人，如果他們想要認識彼此，只有我能充當中間人。所以，如果以我和我妻子為中間點繪製社交圖譜，那麼就比較類似於圖4—4中的右圖。值得注意的是，如果萊西瑪和我在一起工作，或者她不跳舞，或者我們讀的是同一所大學，那麼我們在彼此的社交網絡中就不會扮演現在的角色。

克萊因伯格與貝克斯特倫將其理論稱為「分散理論」，是因為它顯示瞭如果沒有你，你的社交圖譜的分散度。也就是說，如果將你和你的妻子從中心位置移除之後（比如生了第二個寶寶之後淡出社交圈的情況下），你的社交圈子會不會完全垮掉。我更喜歡夫妻之間的「同化」，因為我認為同化能夠更好地體現出社交關係的精髓：同化程度較高的夫妻在多個相互分離的朋友圈中扮演著紐帶般的特殊角色，夫妻二人的共同努力才能讓社交網絡變得更加密切。

同化的力量來自這樣一個事實，即你的配偶往往是唯一一個被你介紹到自己朋友圈的人。你參加工作聚會，她在場；你參加同學聚會，她在場；你和朋友打一天球或打一天遊戲，她也在場。幾乎一年到頭她見證了你參加的每一次公開活動。與此同時，這些同事、同學和球友，雖然內部聯繫非常密切，但彼此之間卻沒有什麼嵌入度，如果不同的群體要發生聯繫，就必須依靠你和你的妻子作為中間人。

夫妻角色的相互同化之所以具有重要意義，能夠用來衡量夫妻關係的牢固程度，是因為對於Facebook上的已婚人士而言，在75%的時間裡，他們的配偶都是其社交網絡同化得最嚴重的人。^[12]更重要的是，如果一對年輕夫妻的同化程度較低，那麼他們離婚的可能性很有可能高出50%。^[13]在最穩定的關係裡，二人深度融入了對方的生活，同化程度比較高。我們還可以考慮一下夫妻之間同化程度較低的情況。在這種情況下，夫妻二人各自擁有相對獨立的社交圈，而且互不參與對方的社交生活，這樣一來，雙方就會把大量的時間和精力投入社交圈，而忽視了對方。這種情況和圖4—4中左圖所示的嵌入度過高的情況是一樣的，都會導致夫妻中的一方忽視另一方。處在一個沒有同化的社交網絡中，獨立的生活很快就會變成私密的生活。這種生活就像圖4—5所表示的這樣。

為了與夫妻角色的同化現象做對比，克萊因伯格與貝克斯特倫還檢驗了其他一些用來評估關係的方法。他們在論文中順帶提到了一個細節，我認為特別有諷刺意味。最初的時候，用來預測夫妻關係的最佳指標跟他們的社交圖譜一點兒關係都沒有。在兩人開始相愛的第一年左

右，最佳的方法是他們瀏覽對方頁面的次數。隨著時間的推移，頁面瀏覽次數會逐漸減少，他們社交網絡的交集會增加，這時，夫妻角色的同化作用才開始逐漸顯現。換句話說，愛上一個人帶來的好奇心、探索欲和視覺刺激，最終會逐漸趨於削弱，夫妻二人會逐漸嵌入對方的社交網絡，而嵌入程度的提升反而會對夫妻關係造成負面影響。

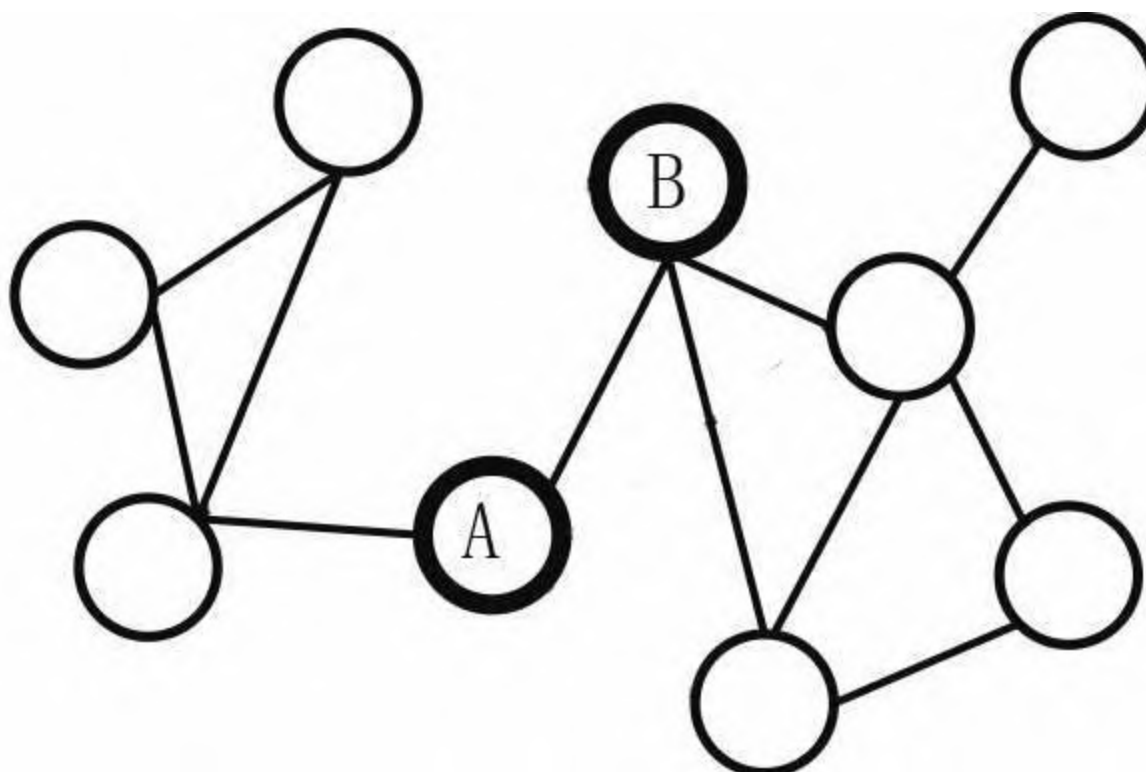


圖4—5 同化程度低的夫妻的社交圖譜

∞

在計算機科學中，有這樣一個理念，即對於自己做出來的產品，你必須是它的用戶。換言之，對於自己向世界推出的網站或軟件，你至少應該有足夠的信心，自己要去用一下。喬納斯·索爾克（Jonas Salk）曾經將自己發明的全新的脊髓灰質炎疫苗注射到自己的身體裡，以檢驗疫苗的效果。同樣，我們也要有勇氣證明自己所做的事情是否正確。程序員將這類事情稱為「吃狗糧」，即邀請公司內部的員工參與到測試中來，提供反饋和建議。員工在第一時間測試這些新技術，可以更快速地發現問題並協助做出改進，藉此提高用戶體驗度。在有些公司裡，這種內部測試是強制性的。微軟公司就是如此。如果你去看一看微軟的員工，就會發現他們都在拿著安裝著微軟視窗操作系統的手機和Surface系列平板電腦，認認真真地驗證其操作性能。[\[14\]](#)

當然，在這裡，沒有人給我們下達這樣的命令，但我在本章伊始仍然拿我自己的數據展開了論述。我之所以這麼做有兩個原因：第一，我需要用鮮活的例子來說明抽象的概念；第二，在本書中，我舉了很多關於其他人的數據，這些數據都具有高度私密性，我願意把這種方式應用到自己身上。

我為你提供了同樣的機會。為了讓你根據本章探討的原理來檢測一下自己的婚姻關係、戀愛關係或不健康的依賴關係，我和我的團隊運用克萊因伯格與貝克斯特倫的理論製作了一款應用程序，你只要在瀏覽器的地址欄輸入dataclism.org/relationshipstest，就能進行測試。

輸入兩個人的Facebook賬號之後，程序不僅會描述出你們社交圖譜的交集以及你們的相互嵌入程度，還會對你同化最多的聯繫人進行排名。當今世界已經發展到了一個我們利用自身數據就可以瞭解自我的程度，我們不必等著米爾格拉姆或歐拉教我們如何認識自己。Facebook與Twitter將我們的數據拿去，讓別人進行學術研究，而我們開發的這款應用程序能夠幫助你自己去解讀這些數據。目前，我們只能利用這些簡單的工具來捕捉和分析自己的活動，但相信不久之後便會出現更好的工具。當你看到一些中層管理人員在電梯裡還在設置他們那具有運動追蹤功能的Fitbit智能手環時，你就知道「量化自我」運動^[15]的確已經開展起來了，而我編寫的應用程序只不過為這項運動進獻綿薄之力而已。

如果你和別人使用我的應用程序，我希望你在對方的聯繫人列表中排在第一位。要記住：適當地刪除一些好友，可以提高你們在同化對方方面的得分。通過這一款小小的應用程序評估自己的社交網絡，可以根據別人在自己生活中的重要性進行排名，讓你一目瞭然地瞭解自己的社交圈，這是非常好的一個辦法，但前提是不要讓它把你的前女友排在你的妻子前面。

[1] 圖4—1和圖4—2中的社交圖譜是由James Dowdell製作的，製圖方法大體類似於Lars Backstrom和Jon Kleinberg在其論文「Romantic Partnerships and the Dispersion of Social Ties: A Network Analysis of Relationship Status on Facebook」中提到的方法。這篇論文發表在2014年2月15~19日於馬里蘭州巴爾的摩市召開的18th ACM Conference on Computer Supported Cooperative Work and Social Computing會議上。該文獲取鏈接如下：delivery.acm.org/10.1145/2540000/2531642/p831-backstrom.pdf。

[2] 我那個樂隊的名稱是Bishop Allen，樂隊的另一個成員是賈斯汀·賴斯（Justin Rice）。你可以在Spotify、BT或iTunes上找到我們的歌曲。如果你感興趣的話，我可以給你推薦以下幾首歌曲：「Like Castanets」，「Click Click Click Click」，「Chinatown Bus」，「Start Again」和「Little Black Ache」。

[3] 雖然我早在大學時期攻讀數學專業時就熟悉歐拉、橋樑問題及其在圖論形成過程中的

作用，但我還是參考了維基百科「哥尼斯堡的七座橋」條目的內容，以期瞭解問題的相關細節和解決方式。

[4] 這七座在歐拉時代非常知名的橋，目前已經消失了四座，其中兩座被炸燬，兩座因修建高速公路而拆除。

[5] 關於圖論在古代和現代的運用情況，下面這個鏈接可以為你提供不錯的參考信息：world.mathigon.org/Graph_Theory。

[6] 對於米爾格拉姆及其成果，我很多年以前就比較熟悉了，就像熟悉歐拉及其成果一樣，但在討論他的「六度分隔理論」實驗時，我依然參考了維基百科的相關內容。

[7] 關於這一點，詳情請參考Johan Ugander等人合寫的「The Anatomy of the Facebook Social Graph」。

[8] 這是喬布斯的創意。我最早是在喬納·萊勒（Jonah Lehrer）所著的《想象》（Imagine）一書中瞭解到這個創意的。See Buzz Feed's 「Inside Steve Jobs' MindBlowing Pixar Campus,」 by Adam B.Vary, for more details.

[9] See 「The Strength of Weak Ties」 by Mark S.Granovetter, American Journal of Sociology 78, no.6 (1973): 1360—80.

[10] 這個概念的原始論文已經被引用了2萬多次。

[11] 「嵌入度」這個概念最早是由格蘭諾弗特（Granovetter）提出來的，但我在本章中，對於嵌入度、人際網絡理論的討論，主要還是來自貝克斯特倫（Backstrom）和克萊因伯格（Kleinberg）的《浪漫關係》（Romantic Partnership）一文。我將他們的啟發方式應用於自己的網絡上，並且考慮到我們的用戶可能不是專家學者，而是一般的民眾，我在應用的過程中進行了一定的簡化。

[12] 貝克斯特倫和克萊因伯格定義了多種具有細微差異的離散模式，我在這裡所列的數字是採用他們所說的「迂迴式離散」（recursive dispersion）方法計算出來的。

[13] 這個觀點援引自貝克斯特倫和克萊因伯格的論文中的一段話：「我們發現，如果無法通過迂迴式離散方法在一個人的人際關係網絡中準確地識別其伴侶，那麼這個人在60天內變為單身（即伴侶關係破裂）的可能性要高出很多。無論這種關係維持多久，這種效應都適用，並且對於維持時間不足12個月的關係特別明顯。在迂迴式離散方法無法正確識別一個人的伴侶的情況下，那麼這個人變為單身的可能性會高出50%。」

[14] 「量化自我」運動，Quantified Self，是指通過科技方式將自己日常生活的各方面，包括物質攝入、身體狀況、體能情況以及其他一些細節記錄下來的一項活動。——譯者注

[15] 並不是所有微軟員工都是這樣。但至少根據我的經驗來看，手機和平板電腦部門的員工只能用自己的產品。內置Windows系統的手機非常罕見，因此，十分引人注目，當你看到它的時候就不會忘記。在這裡，或許可以提一下，我一直是微軟公司開發的Office系列辦公軟件的忠實用戶，本書的所有圖表以及分析藉助了Excel軟件。

第五章 「約會大冒險」：雖敗猶榮

有一個輕博客，叫作「來自地獄的客戶」（Clients from Hell）。在這裡，來自服務行業，尤其是網頁設計行業的網友們晒出了各種奇葩客戶，分享了他們親身經歷的那些莫名其妙的、讓人摸不著頭腦的事情，每隔幾個小時就會湧現出一些新帖子。下面這個故事是一個圖片設計師晒出來的，是一個很有代表性的帖子：

客戶：能給照片加個標題嗎？

設計師：哦，已經有圖片說明了。

客戶：如果讀者忽略了圖片說明，那麼他們仍然會看到標題。

設計師：既要標題，又要圖片說明，這不合常規啊。

客戶：這很有意義啊。那就在圖片說明旁邊加一個標題吧。

我從這個輕博客上引用得最多的一句客戶怨言就是：「我不喜歡這個圖形中的恐龍。它看起來太假了，換成真恐龍的照片。」雖然這個博客主要收集來自平面設計師的帖子，但其流程度卻說明了一個普遍性的真理，即人們對自己的客戶或多或少都抱有厭惡情緒。

我並不是說這種厭惡情緒是針對某一個客戶的，而是從整體上來講，客戶群體就像一群暴民，令人望而生畏。如果有人說他不厭惡客戶，無論是蛋糕店店主，還是董事會會議室裡的首席執行官，那麼可以肯定的是，他在撒謊。雖然他們會說一些諸如「客戶總是正確的」之類的話，但誰都不會真心喜歡如此強勢的客戶。客戶群體之所以會給別人造成嚴重沮喪情緒，部分原因在於他們根本不明白自己真正需要什麼，也無法清楚地將自己的想法表達出來。正如史蒂夫·喬布斯所說的那樣：「人們壓根兒不知道到底想要什麼，直到你將產品放到他們眼前。」但有一句話喬布斯沒有說出來，那就是，如果將一種產品，尤其是科技產品，放到客戶眼前，恐怕會立即跳出幾百人嚷嚷著提各種各樣的奇葩建議。

比如，如果你經營著一家汽車製造企業，而客戶不喜歡你產品的某個部分，那麼大部分情況下，他們都不會直接告訴你，而是通過拒絕購

買的方式間接地告訴你。在歷史上，有些客戶希望汽車的杯座是綠色的，還有一些客戶希望方向盤是正方形的，因為大部分轉彎都是90度的。但福特公司與這些客戶之間一直沒有建立公開的溝通渠道。正是由於缺乏這種公開的溝通渠道，傳統企業才投入大量資金和精力做市場研究，他們必須敏銳地捕捉客戶需求，因為如果他們等到盈利下降才發現客戶需求，再想採取補救措施，恐怕為時已晚。

然而，網站與傳統企業不同。一旦客戶有了新奇的想法，只需發一封電子郵件就能聯繫到網站管理人員。如果用戶不再使用某個服務了，那麼網站也能立即注意到。網站能對用戶使用網站的情況進行事無鉅細的實時跟蹤監測。

如果你在自己最喜歡的網站上，比如谷歌、Facebook、領英（LinkedIn）、YouTube或其他任何網站上看到了一個新服務，並點擊進入的話，那麼網站後臺的計數器就會立即發現，上面的數字就會增加一次，可能一個戴著耳機、吃著薯片的程序員就看到了這個計數器上的數字變化。網站產生的數據量是極為龐大的，甚至會把人逼瘋。谷歌視覺設計團隊的創建者道格拉斯·鮑曼^[1]最後之所以辭職，就是因為當時谷歌公司的視覺藝術設計過程是以龐大的數據為基礎的，而數據量多到了令他難以承受的地步。比如，如果谷歌公司不確定某一個藍色按鈕採用哪一個透明度比較好，就會在內部測評時將41種透明度全部試一遍，以確定哪個效果最好。這會嚴重增加程序員的工作量。德爾斐的阿波羅神廟的門柱上刻有「認識自己」（know thyself）的銘文，而你一打開電腦，很容易湮沒在海洋般浩瀚的互聯網數據中，不僅難以認識自己，就連人類在漫長的歷史長河中總結出來的其他智慧也煙消雲散。

搞清楚客戶對於一輛汽車或一個網頁的期待，是商學院或設計工作室感興趣的事情，而我感興趣的事情是人們為什麼不瞭解自己真實的需求。在社會科學領域內，人們言行不一是很常見的，但由於我是OkCupid網站的運營者，所以有幸觀察到人們在現實中採取的矛盾做法。我之所以花費大量精力去觀察用戶的行為方式，是因為我本人也不知道他們究竟想要什麼。

∞

2013年1月15日，OkCupid網站宣佈將這一天定為「盲目愛情節」（Love Is Blind Day），並且我們非常大膽地決定做一個實驗，把OkCupid上所有用戶的照片都刪除了幾個小時。我們的想法是做一個實驗，併為當天推出的一款手機應用程序造勢。我們的程序員在上午9點

開始刪除圖片，結果與正常的週二相比，我們網站的所有指標都下降了（如圖5—1所示）。

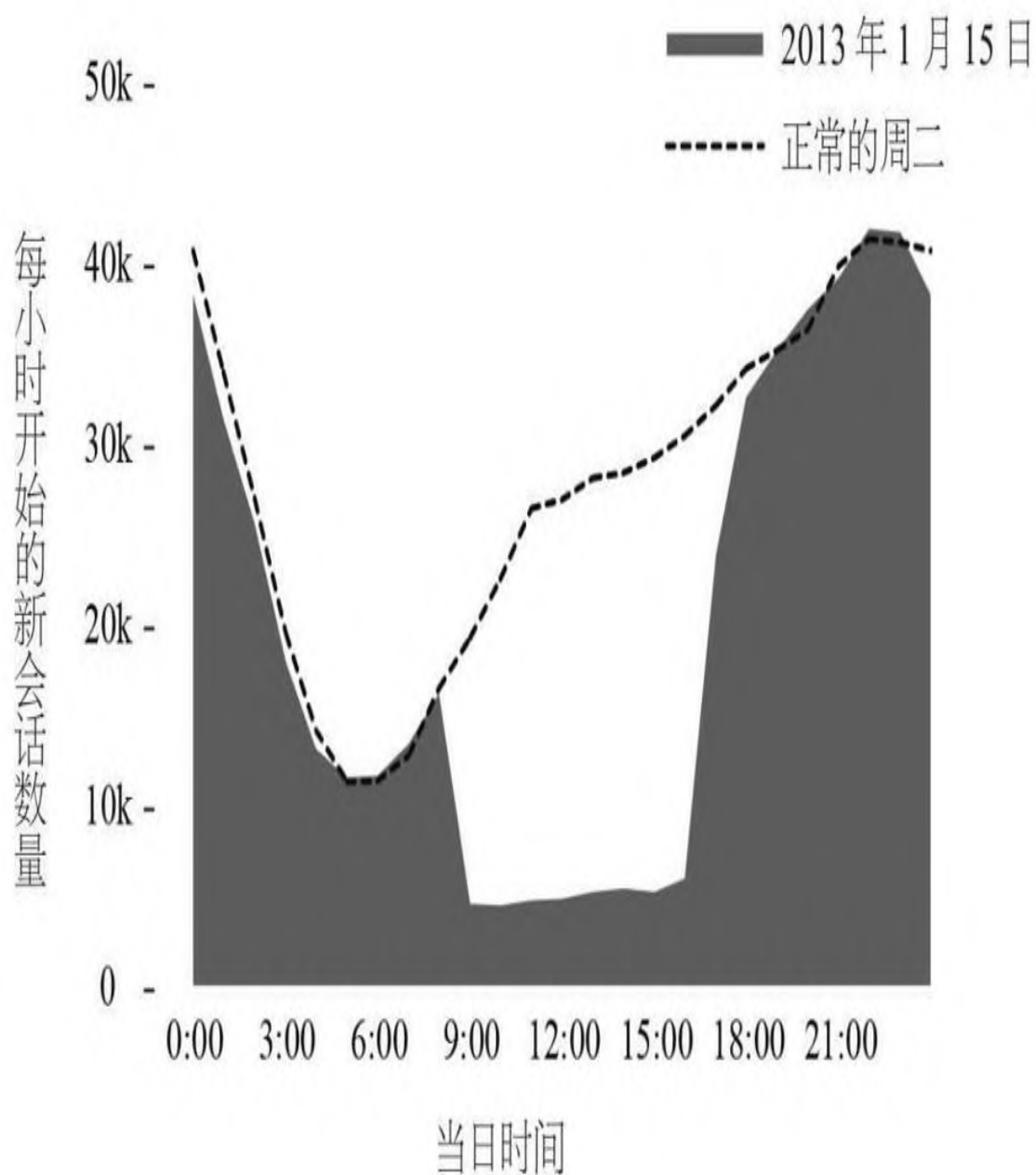


圖5—1 「盲目愛情節」實驗

這種情況真是太令人絕望了，在平時是很罕見的！我們曾經努力宣傳的這款手機應用程序名為「約會大冒險」（Crazy Blind Date）。這款應用程序的目標就是讓大家配對的速度加快，省去在網站上你來我往的時間，直接見面，然後決定是否繼續聯繫。你只需要填寫基本的資料，

上傳一張個人照片，在手機屏幕上敲幾下，選擇你想約會的地方和時間就可以了。這款應用程序會為你自動匹配另一個用戶，並選擇附近的地點和最近的時間，讓你們兩個人見面。這款應用程序會提供一個確認界面，讓你們二人確認是否同意和對方見面，但應用的特點之一是，約會對象所看到的照片是被打碎的，類似一塊被打亂的拼圖，就像圖5—2一樣。雙方在約會之前只能知道對方的名字，卻無法直接溝通。你唯一能做的就是約定的時間與對方見面，並期待著出現最好的結果。



圖5—2 被打亂之後的照片拼圖

你可能已經注意到了，我在說到這款應用程序的時候，用了「曾經」一詞，因為這款應用程序被下載了25萬次之後，最終還是被蘋果公司的應用商店下架了，原因是人們終歸還是堅持先看一看對方的照片。和其他很多應用程序一樣，這款應用程序的創意聽起來很好，但一應用到現實中就行不通了。這款應用程序投入使用的這段時間，我們似乎過了一個長長的「盲目愛情節」。這款應用程序退出數月之後，我們關閉了這項服務。這款應用程序下架之前，大約1萬人通過它同一個自己從來沒有見過、沒有說過話的人分享了啤酒或咖啡。

這些少數比較勇敢的人為世界留下了一組十分珍貴而稀有的數據。「約會大冒險」這款應用程序不僅記錄了A與B兩人曾經真正見過面，還記錄了他們對於彼此的看法，因為每次約會結束之後，這款應用程序就像一個愛八卦的室友一樣，詢問他們約會的情況。因為它的大多數用戶同時也有OkCupid賬戶，所以我們在分析過程中可以運用各種各樣的人口學細節來交叉引用這些數據。在OkCupid網站上，我們只能瞭解到用戶在線互動的情況，而至於他們的約會情況以及約會之後的互評，我們則無從瞭解。然而，通過這款手機應用，我們忽然之間擁有了關於真實約會的數據，並且可以將這類數據與OkCupid網站的數據放在一起分析。這樣一來，你就會發現一些令人驚訝的事情，即兩人的外貌對於約會是否快樂幾乎沒有影響。無論哪一方更好看，或者一方比另一方好看多少，即便一方是絕代佳人，另一方平平庸庸，雙方在約會後給予對方好評的比例一直維持在一個固定水平。所以，外貌上的吸引力並非約會愉快與否的決定性因素。這類來自真實約會的數據顛覆了我在經營交友網站10年間看到的一切。

圖5—3是關於男性的數據。正如我在前面所說的那樣，外貌在約會者的相互評價中只有相對的作用，而沒有絕對的作用。我做這幅圖反映出了這樣一個事實，即人們在約會中的感受如何，在很大程度上取決於他們自身的外貌。圖5—3中間的柱形表示雙方外貌吸引力相差無幾時，男性對於約會的滿意率。在這個柱形兩邊，分別用6個柱形來表示外貌對比情況的變化。如果男女外貌對比情況處於最右側的柱形，即男性外貌遠超過女性，男性對其約會的滿意率竟然高達100%；逐漸向左推移，到第10個柱形時，男方的感覺就大為不同了。雙方外貌吸引力的對比情況對於男性滿意率的影響是很容易預料出來的。我統計了一下約會的次數，這麼多次約會的情況也能證實你的預料。此外，沒有證據表明有人在故意欺騙這個系統，比如，在約會之前就看到了對方的照片，或者當對方出現後，自己由於對對方的外貌不滿意而偷偷地離去。^[2]沒有

證據表明這類現象的存在。在圖5—3中，橫線表示男性的滿意率。

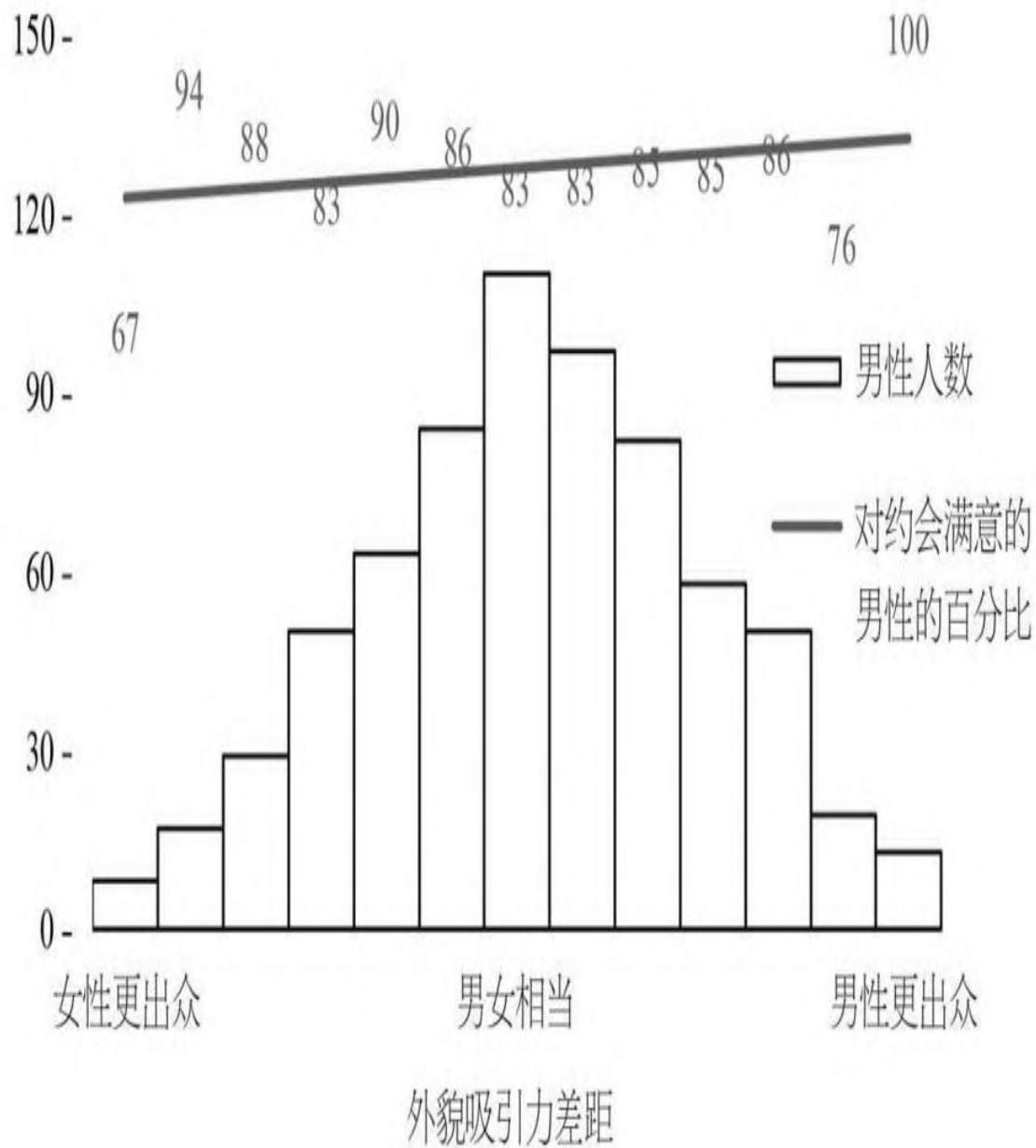


圖5—3 外貌吸引力對男性約會滿意率的影響

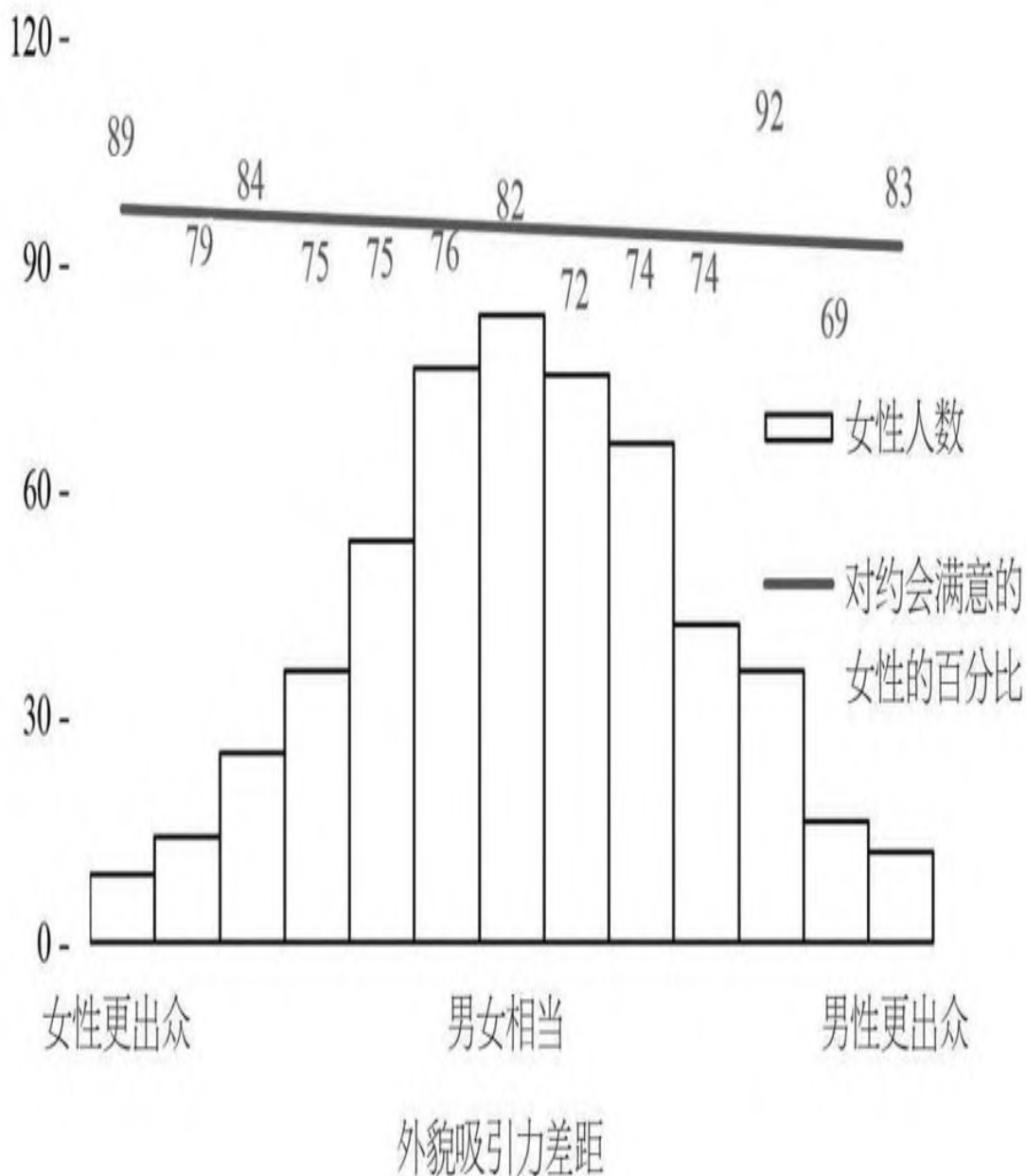


圖5—4 外貌吸引力對女性約會滿意率的影響

圖5—4揭示的是女性對於約會滿意率的影響：

從這款應用程式收集的數據來看，人們對於外貌的重視程度似乎沒有人們普遍所想的那麼高。女性在75%的約會中以及男性在85%的約會中都很愉快。像這種對於外貌的漠視與人們在OkCupid網站上看到的情況恰恰相反。圖5—5對比了同一批女性在真實約會中的滿意率（橫線）以及在OkCupid網站上對於男性信息的回覆率。為了便於比較，我用兩

條實線表示女性的滿意率或回覆率的平均水平，用虛線表示女性滿意率或回覆率與平均水平的差異情況。

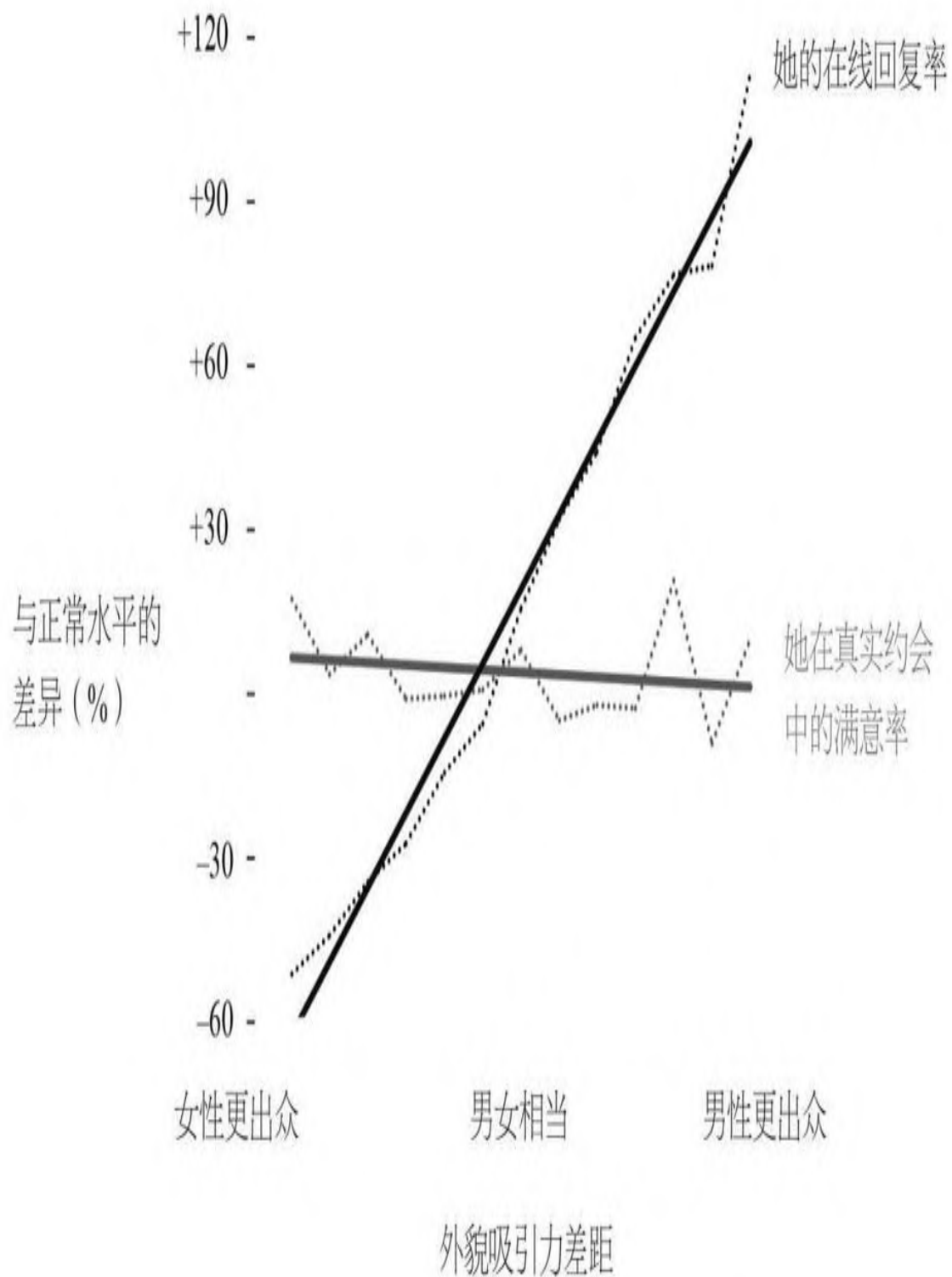


圖5—5 女性對於男性外貌吸引力的反應

男性的圖與此非常類似。需要說明的是，這些數據的背後是同一組男性或女性。黑線表示他們在OkCupid網站上的情況，而灰線表示在「約會大冒險」這款應用程序上的數據。簡而言之，人們在網上進行預先選擇的時候，往往會更加註重一些不太重要的因素，而雙方一旦坐下來，這些因素似乎並不重要了。

這種膚淺的預選無處不在。事實上，這一點有利可圖。你知道退燒止痛藥「泰利諾」（Tylenol）和克羅格公司的對乙酰氨基酚有什麼區別嗎？區別就在包裝盒上。除非你吃藥的時候吞下整個盒子，否則就沒必要支付兩倍的價格去買名稱不同而效果相同的泰利諾。然而，我的桌子上仍然擺著一個紅盒子的泰利諾。

當然，我們非常關注人們身上貼的標籤。在OkCupid網站，有些用戶將自己標榜為民主黨人或共和黨人，但從膚淺的匹配度來看，這類人的匹配度非常低，甚至低於那些自我標榜為新教徒或無神論者的人。我是通過該網站問的諸多匹配問題知道這一點的。這些問題幾乎覆蓋了各個方面，而且平均來講，每個用戶會回答大約300個問題。該網站讓你自己決定自己回答的問題的重要性，並讓你自己描述希望潛在匹配者做出什麼回答，以及最不喜歡潛在匹配者做出什麼回答。雖然網站採取了這一系列控制措施，但涉及政治話題時，系統似乎崩潰了。當通過這些標籤去看一看誰給誰發信息、誰給誰回覆信息以及最終誰和誰約會，你就會發現，不知由於什麼原因，人們會過度關注潛在匹配者對於政治類問題的回答，政治信仰比其他一切信仰都更能影響人們的匹配度。我們在2011年夏季所做的一個實驗印證了這個事實。

在選擇潛在匹配者的時候，人們往往失去自控能力，過於注重很多本來無關緊要的問題，這無異於給對方套上了一些條條框框。比如，有些用戶在設定自己的交友標準時明確指出，自己要找的人必須愛狗，相信不可知論，不抽菸，是個自由主義者，沒生過孩子，等等。但很多人也會問對方一些比較溫和的問題，比如是否喜歡恐怖電影，是否曾經獨自去過另一個國家等。這些問題的指向性是非常明確的，如果對方的某個答案不符合提問者的期待，約會可能就無從談起。如果你在第一次約會時不知道問對方什麼問題，不妨嘗試一下這些方式。在OkCupid網站撮合下建立長期感情的人之間，大約有3/4的情侶或夫妻對於這些問題的回答是一致的，要麼都回答「是」，要麼都回答「不」。人們往往過於注重一些大的、引人注目的因素，比如信仰、政治等，當然還有外

貌。但這些因素並沒有人們所想的那麼重要，有時候，這些因素一點也不重要。

我們的「盲目愛情節」雖然以失敗而告終，卻讓我們真實地看到了人們在信息缺失的情況下做出的行為選擇。通過隱藏照片，其他一切保留不變，我們切切實實地做了一個非同尋常的實驗，這在網站正常運作時期是無法完成的。我們根據之前的數據往往認為照片是最重要的，但在照片被隱藏的7個小時裡，我們的用戶無法根據對方的外貌做出選擇。

在看不到照片的7個小時裡，有一些結果是可以預測的。用戶在發送約會信息時，沒有表現出多少典型的偏見，也沒有對種族問題或外貌問題表現出過度重視。用戶無法對看不到的東西做出評判。在盲目狀態下初次發送的30 333條信息中，最終有8 912條得到了回覆，這個比例比平時上升了40%。此外，還有一件令人震驚的事情。在照片被隱藏起來的7個小時裡展開對話的用戶中，有24%的人在照片回覆之前交換了聯繫方式。這只是在那7個小時裡出現的奇蹟，我們之前的設想只是這個數字的一半。因此，在這段時間內，不僅求交往的信息更容易得到回覆，而且他們與陌生人交換電話號碼和電子郵件的概率也更高。

然而，當我們在那天下午4點恢復照片時，卻出現了一種逆轉效應。之前，兩個人一直處在黑暗中，突然燈亮了起來，導致當時正在聊天和相互瞭解的用戶們突然感到非常驚訝。從我們的數據中就能看出這一點。在恢復照片前後那段時間內，很多正在聊天的人突然停了下來，他們平均交換了4.4條信息，而在正常情況下，他們應該平均交換5.6條信息。照片恢復之後，用戶們交換的聯繫信息也呈現出了類似的減少態勢。

交友網站是為了給單身人士提供相關的工具和信息，幫助他們獲得充滿樂趣的約會，找到伴侶，走進婚姻殿堂等。像身高、政治觀點、照片、自我描述之類的信息都在網上，很容易分類，也很容易搜索。這些是為了幫助用戶做判斷，實現自己的願望，但這樣一來反而容易幫倒忙，即促使用戶在做選擇的過程中過度依賴這些信息。用戶可以將這些信息作為依據，但這些信息不是不可或缺的。

我不禁想起來很多遭到拒絕的用戶。他們之所以被拒絕，是因為一些顯著的因素破壞了他們的好事，而在現實中，一旦兩人坐下來，幾乎沒有人在意這些因素。我在想，互聯網是否像改變了其他諸多事物一樣改變了人們對於愛情的態度。我可以用這樣一句話來表達內心深處對於

交友網站用戶的忠告：在網絡上，你可以得到你想要的，但至於自己真正需要什麼，卻難以發現。

[1] 在科技界，道格拉斯離開谷歌是一個著名的事件。See his own post 「Goodbye, Google」 at stopdesign.com/archive/2009/03/20/goodbye-google.html.

[2] 把「約會大冒險」的照片復原到沒有美化的程度是一件非常容易的事，而且我們也預料到很多人會這麼做。事實也不出我們所料。該應用程序推出後大約一週的時間內，就有幾名黑客寫出了復原照片的程序，只不過那些程序沒有火起來，一方面是因為用起來太複雜，另一方面是因為偶爾還會失靈。我們在分析這款應用程序的產品生命曲線以及生成的相關數據時，這些還原照片的程序都不在考慮範圍之內。本書給出的照片來自圖庫，得到了Getty Images的授權。

第二部分 我們的隔閡從何而來

第六章 混淆變量

如果你站在紐約第五大道和58街交叉路口的西南角，拿著一個寫字板，觀察一會兒來來往往的行人，那麼你很快就能得出這樣的結論：大多數紐約人都漂亮和苗條，最重要的是富有。人們衣服上的每一條線、每一個金屬釦眼、每一個折縫都閃耀著金錢的光芒。當然，許多紐約人是富裕的，但這並不是故事的全部。要知道，你所站的位置是紐約最大奢侈品店之一波道夫·古德曼（Bergdorf Goodman）外面，這個因素就構成了一個「混淆變量」。

「混淆變量」這個術語指的是在分析過程中無法控制的變量，有時也稱為「額外變量」。這些變量可能會影響到自變量與因變量，但因為沒有受到控制，所以其影響是個未知數。如同在觀察人群時不能只在紐約上東區的繁華地段一樣，在收集和處理數據過程中，也不能只考慮理想情況，要確保將混淆變量考慮進去。這是一件耗費腦力的事情，需要花很多時間去規劃。你認為自己在分析和猜測過程中似乎已經考慮到了每一個變量以及每一種可能性，可以根據自己的興趣對這些數據進行研究，但一句老生常談的話說得不錯，自由需要永恆的警惕。

在這裡，我得承認一個事實。到目前為止，你在本書看到的每一個評價、評分、圖、表、比率、總數以及「約會大冒險」上的每一個約會結果都是關於白人的。我不得不這樣做，因為如果你想看一看美國的兩名陌生人在浪漫環境中的行為方式，種族是最終的混淆因素。為了確保這些數據符合我對性吸引力或性愛的想法，我在討論過程中需要暫時把其他種族略去不談。

對於一個美國人而言，掩蓋和忽視種族問題是一種與生俱來的條件反射。雖然我收集的這些數據有失公允，但我的做法也是自然而然的。美國與種族問題具有一種特殊的關係，長期以來，美國在種族問題上一直都在做表面文章，美國某些學校或機構雖然會接受少量黑人以表示種

族平等，但這種做法是蓄意為之的，並非真心認識到了種族平等的意義。此外，美國一些白人在種族問題上長期開展所謂的科學研究，力圖證明白種人的優越性，但這是一種令人遺憾的偽科學。在這種情況下，要對種族問題進行量化分析尤其困難。這種困難並不是說我們不掌握數據；相反，我們有很多這類數據，有的是關於某一個特定領域的。^[1]如果說我比較喜歡的數據是人與人之間的數據，那麼那些業已存在的數據是關於人與物的數據，比如，某一種群人在就業率、學習能力測驗、刑事司法體系、癌症等方面的數據。雖然這類研究能夠幫助我們發現不平等現象，偶爾也能幫助我們解決不平等問題，但數據本身有不完善之處，因為你沒有考慮到誰在做招聘、教學、警察和預防保健，沒有考慮到誰造成了這些統計結果。因此，你最後就得到了這樣的結論：同一宗罪的黑人被告比白人被告入獄的概率高出30%。^[2]這個結論的消極語氣說明了一切。誰導致了司法的錯誤呢？從這個句子的句法本身來看雖然找不出什麼線索，但實際上，我也能猜出個大概。很少有研究會通過這些表面現象，從「我們與他們」的角度去深入地研究種族關係。

在我收集的數據中，每一字節數據的背後都有兩個人，即主動者與被動者，而且我以平等的態度對待每一個人，這的確是一個新現象。本書的英文書名Dataclysm並不僅僅是一個不同字母組合而成的聰明的雙關語，clysm表示龐大的數據如同洪流一般。這些數據是全面的，而不是在某一個特定的時間對某一個特定人群的統計，所以我們才能完整地分析人類經歷。

在我這些數據出現之前，公共生活中量化程度最高的一個領域就是體育領域。在這個領域內，關於每一次賽事，你都可以獲得非常準確的實時數據，甚至能具體到個人，你可以根據自己的需要進行分割和重組。不過，雖然體育領域內有這麼詳細的數據，人們在討論種族問題時，只是一味地爭論來爭論去，卻幾乎沒有人做過任何深入的分析，這的確令人驚訝。在21世紀頭10年一直存在的關於「黑人四分衛」的爭議就說明了這個問題。多年來一直存在這樣一類週期性的新聞報道，意思是非洲裔美國人在選秀過程中或某個引人關注的比賽中可能會佔優勢，但隨後必然有人會說黑人在美國國家橄欖球聯盟（NFL）這種高水平的比賽中，成功率卻比較低。持有這種觀點的人給出的理由往往是黑人的智商不高，而四分衛需要比較高的智商。這種觀點一旦見諸報端，就會激起大量的反駁和爭論，認為這是典型的種族歧視，是心胸狹隘的先入之見。儘管黑人智商問題對黑人四分衛成功率的影響激起了一波又一波的評論、批評和抗議，但當我分析谷歌的9.7萬條關於「黑人四分衛」

的搜索結果時，卻發現只有一篇文章真正地統計了人們對黑人四分衛和白人四分衛的評分情況。^[3]結果表明，黑人四分衛與白人四分衛的平均得分均為81.55分。從數字角度比較黑人四分衛和白人四分衛符合每個人的直覺。但這些數據並不一定是完美無缺的，很多時候，本應該用數據說話，但人們只會爭論，只會關注運動員的逸聞趣事，而且每個人都認為自己的分析是正確的。事實上，雖然眾說紛紜，但81.55分的平均分證明黑人四分衛和白人四分衛處在同一個水平，之前相互爭執不下的兩派人中間肯定有一派是錯誤的。順便說一下，這篇統計了兩類四分衛平均得分情況的文章並不是發表於某個名不見經傳的博客上，而是發表於《今日美國》（*USA Today*）旗下的博客《遙遙領先》（*The Big Lead*）上，而Twitter和Facebook上卻找不到一個人對這篇文章點贊，這表明美國人並不想知道這個數字背後的故事。

在這類情況下，我們似乎缺少通過統計數據的鏡頭來審視種族問題的主觀意願。之前，在公共生活中，大多數領域並不像橄欖球那樣實現了高度量化，但這種情況正在迅速改變。

在OkCupid上，要比較一個黑人和一個白人，或來自任何種族的兩個人，最簡單的方法之一就是看他們的「匹配度」（match percentage）。這個網站利用這個指數來表示兩個人容納對方的程度。它向用戶提出一系列問題，用戶給出自己的答案，然後一個算法會預測兩人能在多大程度上一起愉快地喝啤酒或吃晚餐。與OkCupid網站上的其他特徵不同的是，網站在計算匹配度時，不涉及頭像等可視化的元素。兩個人之間的匹配度只反映了他們的「內在自我」或者說「內心世界」的構成，包括他們的信仰、需求、要求和興趣，並不涉及兩個人的外貌特徵。從這個匹配度來判斷，OkCupid網站上的四大種族（黑人、拉丁裔、亞裔、白人）是非常相似的。^[4]事實上，種族對匹配度的影響低於宗教、政治或教育的影響。在用戶認為對匹配度具有重要影響的因素中，與種族最相近的因素是星座，而星座對匹配度根本沒有任何影響。對於不會對人和物進行分門別類的電腦而言，「亞裔」「黑人」「白人」與「白羊座」「處女座」「摩羯座」是沒有區別的。^[5]

但種族中立論只是停留在理論層面，因為計算機的算法系統是沒有種族偏見的，而如果考慮到用戶自己的觀點，情況就會發生變化。我們在前面所講的匹配度只是計算機根據用戶對某些問題的回答計算出來的，用戶並沒有看到彼此的頭像。如果讓用戶看到完整的資料，頭像佔據了很大一部分頁面，那麼從種族來看，OkCupid的用戶對彼此的評分

就呈現出了表6—1所示的情況。

表6—1 OkCupid網站上男性對女性的平均評分

		她的种族			
		亚裔	黑人	拉丁裔	白人
他的种族	亚裔	3.16	1.97	2.74	2.85
	黑人	3.40	3.31	3.43	3.23
	拉丁裔	3.13	2.24	3.37	3.19
	白人	2.91	2.04	2.82	2.98

我在上面給出的評分是原始數據，沒有經過任何修飾。你肯定也知道OkCupid網站在用戶評分制度上採用的是5分制。為了更容易看出這些數據的變化趨勢，我將用另一種方式來呈現。^[6]在表6—1中，我先計算出每一行數據的平均值，並將其視為「正常值」，表示某個種族的男性對女性的平均偏好程度，然後計算出每一個數據與平均值相差的百分比。如果一個數據高於平均值，則用加號表示，反之則用減號表示。表6—2與表6—1背後的信息是相同的，只是表現方式不同而已。在表6—2中，你可以看到，亞裔男性認為亞裔女性的漂亮程度比女性平均水平高出18%，而黑人男性則認為亞裔女性的漂亮程度僅僅比平均水平高出2%。

表6—2 OkCupid網站上男性對女性的平均評分與正常值的差異

		她的种族			
他的种族		亚裔	黑人	拉丁裔	白人
	亚裔	+18%	-27%	+2%	+7%
	黑人	+2%	-1%	+3%	-4%
	拉丁裔	+5%	-25%	+13%	+7%
	白人	+8%	-24%	+5%	+11%

稍後，我還會呈現來自其他交友網站的數據，但不再給出原始數據，而是直接列出平均評分與正常值的差異。通過觀察表6—1和表6—2，我們很容易得出這樣一個結論：男性往往更喜歡自己種族的女性，而且男性普遍不太喜歡黑人女性。用戶之間相互發送的信息數量與這些評分數據具有密切聯繫，因此也呈現出了同樣的態勢。^[7]

為了表明這些評分的變動趨勢具有可靠的數據基礎，我在圖6—1中呈現了每個用戶的原始評分數據，它能告訴你一個種族大部分女性得到的評分處於哪個位置。你可以看到絕大多數黑人女性的得分都低於其他三個種族女性的得分。黑人女性的最高得分只大致相當於其他三個種族女性得分的中位數。

從數學角度來看，只要你是一位黑人，那麼你的評分肯定會大打折扣，即便你獲得的評分處於最頂端，基本上也比其他種族的人獲得的評分低出一大截。如果你進行更加深入的分析，看一看投票的用戶都有哪些人，就會發現這種情況是由無數個白人、拉丁裔和亞裔的投票用戶導致的，而不是由一小撮種族主義者導致的。^[8]

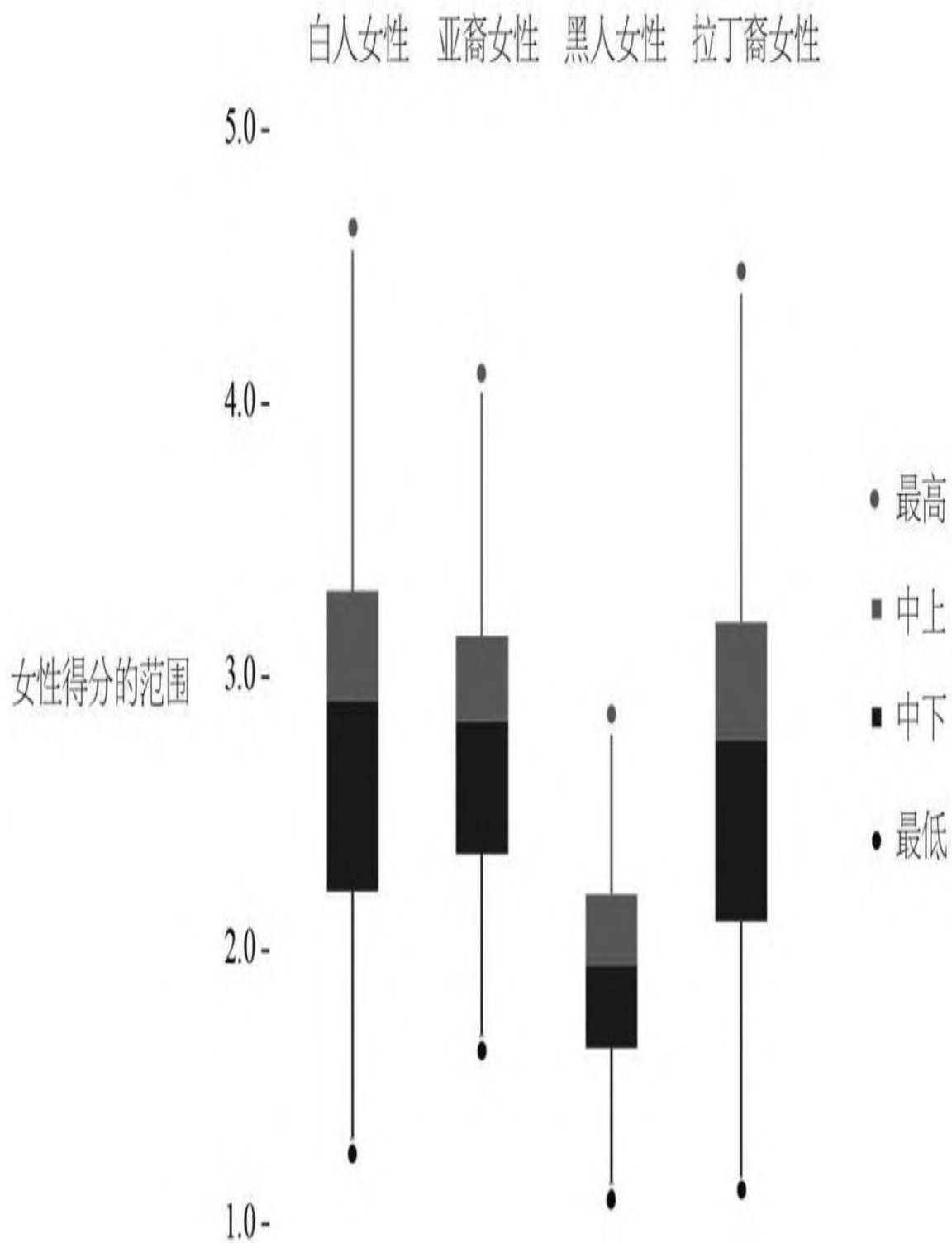


圖6-1 各種族女性得分

無論這一點看起來多麼令人驚訝，它都是一個真實的數據集，反映的是一群人的想法。因此，到這時為止，我們就不得不暫停一下，回答一個你之前可能就會提出的問題，即鑑於我在本書中如此依賴OkCupid網站的數據，那麼這些用戶究竟是什麼人呢？

按照最膚淺的方式來看，OkCupid網站用戶的情況反映了互聯網用戶的普遍情況，當然，還要補充一點，即這個網站上的用戶都是單身，宗教色彩不太濃厚，平均年齡為29歲，低於美國所有互聯網用戶的平均年齡。其種族構成可能你已經猜到了。表6—3列出了該網站用戶與美國所有互聯網用戶的種族構成情況。後面這個數據來自媒體評測和網絡分析公司Quantcast，該公司與Nielsen公司的業務具有一定的相似性。^[9]

表6—3 種族構成

	OkCupid用戶	美国互联网用戶
亞裔	6%	4%
黑人	7%	9%
拉丁裔	8%	9%
白人	80%	78%

如果我們從人口學層面上進一步分析，就會發現與美國互聯網用戶的整體情況相比，OkCupid用戶中的城市居民所佔比例更高、受教育程度比較高，而且思想更加進步與開明。該網站的最大市場是紐約、舊金山、洛杉磯、波士頓和西雅圖，85%的用戶讀過大學，自我標榜的自由主義者是自我標榜的保守主義者的兩倍多。整個網站都瀰漫著一種開放、開明的精神氣質。在回答「你會考慮同一個對某個種族表達出嚴重歧視的人約會嗎」這個問題時，84%的用戶給出的答案是「不會」（共有三個選項：會、不會、不一定）。這些用戶的選擇都是自發的，而非蓄意為之，因此，這個數字的确令人欣喜。根據之前的數據，這就意味

著OkCupid網站上84%的用戶不會考慮同那些存在種族歧視的用戶進行約會。

從本質上來講，OkCupid用戶的種族主義意識並不濃厚。我曾經向很多人提起過這一點。這些人和我一樣也在大城市過著美好的生活，他們認為自己的觀點和品位是非常開明的，他們晚上會通過一兩杯紅酒或瀏覽Facebook放鬆身心，他們具有進步和正義的思想。通過研究數據，我指出黑人女性以及黑人男性會相對受到忽略，而如果一個人只要擁有白人血統，就會使自己變得更有魅力。我這麼講，是在描述我們的世界，這既是我的世界，也是你的世界。如果你在讀一本關於大數據發展前景的大眾科學書籍，請放心，你肯定是那些數據的一部分。

我們再回頭看看前面提到的那個匹配問題：「你會考慮同一個對某個種族表達出嚴重歧視的人約會嗎？」這個問題是由OkCupid的一個用戶提出來的，已經有接近200萬人給出了自己的回答。實際上，我覺得這個問題無異於在問：「你會和一個種族主義者約會嗎？」我曾經認為這就是這個問題的真實意圖，然而，事實並非如此。這個問題的提問者比我更早地理解網絡數據的微妙特徵，他這個問題並非針對真正的種族主義者，很多人只是在交友網站上無意識地或一時衝動地表達出對某個種族的歧視傾向。畢竟，如果不是在網絡上，有一些衝動就會受到抑制。在某種程度上，用戶既會判斷別人，也會受到別人的判斷，每個人在網絡中面臨的環境都不同於日常生活的環境，網站也不會將你的家庭和你聯繫在一起，也不會將你的動向轉發給你的朋友。交友網站的遊戲規則是：它給你推薦一些人，你可以喜歡，也可以不喜歡；你可以跟他們聊天，也可以拒絕跟他們聊天。僅此而已。在一個數字化的網絡世界中，在線約會是在一種相對隱祕的環境中進行的，只涉及你和你選擇的人，你所做的事情是私密的。

事實上，你的朋友通常不知道你註冊了賬號，更不知道你做了什麼。因此，人們在網絡世界中可以相對擺脫現實世界的壓力，根據自己的態度和慾望去行事。

在外行看來，Facebook這個「社交網絡」是收集數據的必然來源。其中的原因是顯而易見的：Facebook的用戶群體非常龐大，幾乎覆蓋了世界各地能接入互聯網的人群；換句話講，無論你想找什麼類型的數據樣本，都能輕易地找到一組頗具代表性的用戶。此外，Facebook功能強大，數據多樣化，其中包括你的高中同學是誰、你剛剛在Spotify音樂服務平臺上聽過什麼歌曲以及你的父母住在哪裡等。

然而，豐富的數據利弊並存。Facebook的一個弊端就是，你在上面很少遇到陌生人。根據該網站的設計原理，你在上面接觸的都是你已經認識的人和你已經決定加為好友的人。畢竟，上面的人都是你的朋友。你經常聽到很多人說自己有黑人朋友，而且既然是朋友，那麼你對待朋友的方式與對待其他人的方式是不同的，這樣一來，必然會影響到Facebook上關於種族的數據的真實性。事實上，你和你朋友的關係是在網絡之外的現實世界中形成的。

此外，在眾多朋友的注視下，人們會感到壓抑。在Facebook上，你的一舉一動都會引起他人的關注。正是由於這個原因，很多約會應用軟件所做的第一件事情就是讓你擺脫毫無隱私可言的環境，營造一個私密的氛圍。很久之前，我們曾經嘗試過在OkCupid網站上添加一些類似於朋友圈的社交功能，結果以失敗而告終；當交友網站Match.com嘗試添加類似功能時，也失敗了。原因之一必然是用戶不希望朋友們關注自己的在線約會。我想象了這樣一個情景：當我們在某家飯館進行一次很有希望成功的約會時，附近的座位上卻坐著我們的兩個老朋友，那麼這種情況就類似於Facebook創建的網絡社區，失去了私密感。

這種真實的人際關係，使得人們在社交網絡上表達的觀點可能不符合自己的真實想法，從而影響到了社交網絡數據的真實性。在種族之類的問題上，人們——至少是那些正派的人——在公開發表言論時可能會感受到壓力，從而不得不遵循某種適當的言論模式，而注重私密性的交友網站則能提供一組與眾不同的數據。因為在交友網站上，用戶都是陌生人，他們會根據自己的真實願望告訴你自己喜歡誰以及不喜歡誰。^[10]

因此，接下來，我們將來自OkCupid網站和其他交友網站的數據對比一下，看看能得出什麼明顯的結論。看一看其他用戶通過其他界面產生的數據，會讓我們更好地瞭解真實情況。下面我們看到的數據來自OkCupid、DateHookup和Match.com這三個交友網站。僅僅在2013年，這三個網站的新註冊用戶就達到了2000萬人左右。從細節上來看，表6—4中的數字可能存在一些差異，但記住，這些數字反映了不同的人在不同軟件上的行為，這些數據的背後卻隱藏著相同的行為模式。從總體感覺的「方向」來看，也就是從「喜歡」或「不喜歡」來看，這些數據幾乎是相同的。

表6—4 三個網站不同族群男性對女性的評價

		她的种族			
他的种族	OkCupid	亚裔	黑人	拉丁裔	白人
	亚裔	+18%	-27%	+2%	+7%
	黑人	+2%	-1%	+3%	-4%
	拉丁裔	+5%	-25%	+13%	+7%
	白人	+8%	-24%	+5%	+11%

		她的种族			
他的种族	Match.com	亚裔	黑人	拉丁裔	白人
	亚裔	+50%	-68%	-14%	+31%
	黑人	+9%	-13%	+8%	-3%
	拉丁裔	+4%	-67%	+33%	+29%
	白人	+13%	-68%	+8%	+47%

		她的种族			
他的种族	DateHookup	亚裔	黑人	拉丁裔	白人
	亚裔	+11%	-24%	+9%	+4%
	黑人	+7%	-9%	+9%	-7%
	拉丁裔	+12%	-27%	+10%	+6%
	白人	+18%	-30%	+6%	+5%

你可能知道Match.com這個網站，這是美國近20年來最受歡迎的交友網站。該網站在全國性的電視節目上大做廣告，因此，用戶覆蓋整個美國，無論你希望找到什麼樣的人，都能從這個網站上找到。

DateHookup是免費的交友網站，擁有幾百萬會員，深受隨意約會者的歡迎。它的用戶群裡，黑人的比例不到20%，拉丁裔的比例約為13%，是這三個網站中最多樣化的一個。如果根據其用戶的族群構成情況將這些網站與美國城市做類比，我認為DateHookup類似於亞特蘭大或休斯敦，OkCupid類似於波特蘭，而Match.com類似於達拉斯。但是如你所見，在這三個網站中，男性對女性的評分體現了同樣的態勢。

反向的評分，也就是女性對男性的評分，雖然在不同網站上呈現了不盡相同的態勢，但總體來看，它們仍然是非常相似的（見表6—5）。

表6—5 三個網站不同族群女性對男性的評價

		他的种族			
她的种族	OkCupid	亚裔	黑人	拉丁裔	白人
	亚裔	+19%	-38%	-15%	+35%
	黑人	-34%	+52%	-17%	-1%
	拉丁裔	-35%	-20%	+19%	+37%
	白人	-26%	-19%	-1%	+46%

		他的种族			
她的种族	Match.com	亚裔	黑人	拉丁裔	白人
	亚裔	+3%	-7%	-5%	+9%
	黑人	-9%	+10%	-1%	+0%
	拉丁裔	-8%	-6%	+6%	+8%
	白人	-7%	-5%	0%	+12%

		他的种族			
她的种族	DateHookup	亚裔	黑人	拉丁裔	白人
	亚裔	-	-34%	+14%	+20%
	黑人	+9%	+25%	-12%	-22%
	拉丁裔	-18%	-14%	+21%	+10%
	白人	-12%	-25%	+7%	+31%

這些數據顯示了兩個消極趨勢和兩個積極趨勢。從消積的角度來看，不僅黑人再次遭到了其他族裔用戶的漠視，而且亞裔男性也遭到了其他族裔女性用戶的漠視。從積極的一面來看，女性顯然更加喜歡自己族裔的男性，女性對族裔的忠誠度高於男性，但除了本族裔男性，她們明顯更加青睞白人男性。

OkCupid網站為我們深入研究種族的「等級」提供了另外一種方法，而研究結果印證了白人比較容易受到其他族裔的歡迎。由於OkCupid的用戶可以將自己的族裔選擇為混血，這就為我們研究種族融合創造了一個幾乎像實驗室那樣理想的條件。比如，有些用戶將自己的族裔選擇為「亞裔」，而有些用戶選擇為「亞裔」和「白人」的混血。將這兩類人獲得的評分做個對比，就能在一定程度上看到增加「白人」血統之後會給一個人帶來什麼。結果表明：變化相當大。當你有白人血統時，評分會全面提高。我在這裡給出了完整數據（見表6—6）。這個表格顯得很大，數據龐雜，但值得深入探究。

在右列的數據中你可以看到，用戶的種族構成中增加了白人血統之後得分改善。最大的改觀是之前一直遭到其他種族漠視的黑人男性、黑人女性以及亞裔男性，在增加了白人血統之後，一改頹勢。

不幸的是，沒有足夠多的人選擇「黑人+拉丁裔」混血或「亞裔+黑人」混血，無法進一步佐證我們的觀點。但這是一個有趣的現象，能幫助我們窺探到我們對不同種族的看法：

表6—6 不同族群女性對混血男性的評價

他的种族		她的种族		
	男性评价女性	拉丁裔	“拉丁裔+白人”混血	百分比变化
	亚裔	2.7	2.8	+4
	黑人	3.4	3.4	-2
	拉丁裔	3.4	3.4	+1
	白人	2.8	3.0	+7
		黑人	“黑人+白人”混血	
	亚裔	2.0	2.3	+19
	黑人	3.3	3.5	+5
	拉丁裔	2.2	2.9	+28
	白人	2.0	2.5	+24
		亚裔	“亚裔+白人”混血	
	亚裔	3.2	3.0	-5
	黑人	3.4	3.6	+5
	拉丁裔	3.1	3.3	+5
	白人	2.9	3.0	+2

表6—7 不同族群男性對混血女性的評價

她的种族		他的种族		
	女性评价男性	拉丁裔	“拉丁裔+白人”混血	百分比变化
	亚裔	1.7	1.8	+7
	黑人	2.0	2.4	+18
	拉丁裔	2.1	2.2	+8
	白人	1.8	2.1	+15
		黑人	“黑人+白人”混血	
	亚裔	2.5	1.6	+6
	黑人	2.7	2.6	-4
	拉丁裔	1.7	1.9	+17
	白人	1.6	2.0	+26
		亚裔	“亚裔+白人”混血	
	亚裔	2.0	2.1	+4
	黑人	1.8	2.7	+48
	拉丁裔	1.5	2.2	+44
	白人	1.5	2.0	+32

表6—6中列出的數據都是交友網站上的評分，而約會數據本質上反映了人們對彼此的第一印象，是乍看之下給出的數據，因為用戶在接吻之前需要了解對方，至少得有一丁點兒的瞭解。任何兩個人，在走到一起之前肯定要先想一想：哦，我看到了什麼？我看到的是誰？上面這些數據描述的是人們遇見陌生人時的激動。在真正瞭解對方之前你是否喜歡一個人，取決於你瞬間爆發的判斷力和本能直覺，取決於對方是否能給你帶來讓你眼前一亮的感覺。OkCupid用戶用自己的話表達了這一點：[\[11\]](#)

後來，有一天，我瀏覽網站每天為我推薦的約會對象時發現了他。我立即點開了他的簡歷……關於他的一些事情總會讓我微笑。

——貝拉評價帕特里克

有一天，我正在瀏覽網站每天為我推薦的約會對象，看到這個女孩。第一眼看上去我就發現她很有魅力，然後，這一切就開始了。

——丹評價簡恩

如果存在一見鍾情，那麼也存在一見生厭，不是嗎？有的人一看到陌生人就會本能地、無意識地退縮，難道不也是因為第一次見面產生的激動情緒嗎？只不過這種激動不是積極的情緒，而是消極的情緒。

下面，我們看一下某人的原話：

在這個國家，很少有非洲裔美國男子沒有在商場購物被跟蹤過的經歷，也包括我；在這個國家，很少有非洲裔美國男子在大街上行走時沒有聽到過旁人忙不迭鎖上車門的經歷。我也遇到過這樣的事，至少在我當參議員之前遇到過……很少有非洲裔美國人沒有看見過在坐電梯時一個婦女緊張地捂著自己的錢包並且屏住呼吸直到走出電梯為止的經歷。[\[12\]](#)

——巴拉克·奧巴馬，2013年7月19日

這些現象的背後都是人的本能在作怪，我們根據少量的信息就能推斷出這一點。這種本能不僅會影響到人們的浪漫關係，還會影響到你會把自己的公寓租給誰，影響到你是否批准一項貸款，當然，也會影響到警察的工作。因為警察往往需要在一瞬間做出決定，而他們又沒有足夠的時間去仔細分析情況，不得不依靠本能。甚至在深思熟慮的情況下，第一印象也扮演著重要角色。比如，2004年，哈佛大學經濟學家森德希爾·穆萊納桑（Sendhil Mullainathan）和芝加哥大學經濟學家瑪麗安·貝

特朗（Marianne Bertrand）曾經做過這樣一次實驗。^[13]他們拿兩份完全相同的簡歷，把一份申請人的名字換成白人常用的名字（Emily、Greg之類的），把另一份的名字換成明顯的黑人名字（Lakisha、Jamal之類的），沒有照片。除了名字隱含的種族信息，其餘部分兩份簡歷完全相同，然後他們把兩份簡歷隨機投給波士頓和芝加哥的大量招聘單位，然後看兩份簡歷獲得的面試機會。結果，白人名字比黑人名字獲得面試機會的概率高出50%，但性別、地域、工作類型對這一差距的影響不大。兩份完全相同的簡歷，僅僅是把名字改了一下，就能有如此明顯的差異。這說明雖然很多公司將自己標榜為「給予均等機會的僱主」，宣稱在招工方面不搞性別、膚色、種族歧視，但實際上他們和其他人一樣也存在種族歧視傾向。

這種具有諷刺意味的現象表明，我們不僅需要以大數據為基礎進行宏觀研究，還需要基於個體層面的微觀研究。當你讀到上一段提到的研究結果，並且看到賈馬爾（Jamul）沒有找到工作，那麼你很可能對那些招聘者搖頭嘆息，表達自己的不滿，因為正是這些招聘者才導致他走了黴運。而我們在本章看到的數據表明，種族主義並不是某些人特有的問題，而是一個普遍性的問題。我們在前面看到具有不同閱歷的用戶在三個不同的交友網站上表現出了相同的行為模式，其中有男性，也有女性；有免費網站，也有僅供付費用戶使用的網站；既有對待約會非常嚴肅的人，也有非常隨意的人；既有城市化程度較高的用戶，也有更多的「主流」用戶。總體來說，這些研究代表了美國一大部分年輕人的想法，而且研究出來的數據一致地表明，非洲裔美國人遭到了其他種族的忽視。之所以出現這個問題，並不是因為少數黑人用戶長得醜，也不是因為少數頑固不化的種族主義者導致了研究結果的歪曲，而是具有牢固的現實基礎——絕大多數美國人都在潛意識裡存在種族歧視觀念。

與以往不同的是，現在在公開場合公開發表種族主義言論再也無法為社會所接受。為了應對這種壓力，有一部分人便悄然發洩：如果我再也不能對著學校的孩子們大吼大叫來表達痛恨，那好吧，我就對著電視吼出來。不過，一般美國人不至於扭曲到這種地步。我們大多數人——實際上，幾乎所有人——都意識到了種族主義的錯誤性，不過我們所做的很多決定都暗含著種族主義色彩。^[14]心理學家用「圖式」

（schema）一詞來指代這種內在的信仰模式。圖式是人腦中已有的知識經驗的網絡，也表示人們對特定概念、事物或事件的認知結構，它影響對相關信息的處理過程。大多數人的圖式與自己的世界觀仍然存在一定的差距。在數以百計的日常行為中，我們可能不會刻意懷有種族主義的

意圖或意識，但其實這些行為反映了一種廣泛存在的文化。正如我們所看到的那樣，我們的行為模式如此根深蒂固，以致近些年新加入美國社會的亞裔和拉丁裔也養成了種族歧視的習慣。

談到這些種族主義的行為模式，個人在某種程度上是無可指責的。事實上，對於黑人在交友網站上獲得的關注遠遠少於其他種族，我也深感意外。但我不能因為一個人不想跟另一個人約會就對其橫加指責，用戶的決定也幾乎沒有任何惡意。在交友網站上評價一個人是一件看似微不足道的小事，人們的決定往往是在一瞬間做出的。當你隨意瀏覽網頁時，也許12張頭像中只有一個是黑人，看著這個人的時候，你可能會產生任何一種感受；你看到一個白人用戶的頭像時，也可能產生任何一種感受。也就是說，你的感受是一直處在變化中的。每個人都有權對他人形成自己的看法，有權喜歡一個人，也有權拒絕一個人。那麼，如果你在某個時刻不喜歡某個人怎麼辦呢？這個時候，我們不應該將對某個人的厭惡轉移到他所代表的那一類人身上。事實上，將一個人視為一個個體，而不將其等同於一類人，是朝著正確方向邁出的正確一步。社會一直在進步，只是我們從整體角度來看時，才發現社會仍然存在一些歷史上存在的不足之處。打個比方，如果把我們的社會比作賭場，那麼黑人就相當於賭場中落魄的輸家。之所以有贏家，也有輸家，不是莊家導致的，不是賭客導致的，更不是各種欺詐花招導致的，而是因為有輸有贏是一種既定的遊戲規則。賭場如此，人類社會也是如此。

社會學教授奧塞奇·奧巴索契（Osagie K. Obasogie）最近開展了一系列頗有創意的研究。他採訪了一些出生時雙目失明的人，結果發現他們對待種族問題的態度和那些視力正常的人是一樣的。他選取的樣本量相對較小，只有106人，卻印證了OkCupid網站上的數據。在他舉的許多例子中，一位年輕的盲人一開始對某次約會非常高興，後來撫摸對方頭髮的感覺或者一個陌生人的耳語使其意識到對方是黑人，最後導致那次約會戛然而止。

奧巴索契認為，盲人對於種族的態度反映了一種文化背景下個體在其一生中對另一種文化的吸收與認同程度，而不是反映了什麼視覺現實。從他的數據中我們似乎只能得出這一個結論，而不可能總結出截然相反的觀點。此外，他指出，在傳統文化的影響下，不同種族之間最容易出現失調的方面就是與性有關的問題。正如他在接受《波士頓環球報》採訪時所說的那樣，即便他調查的盲人群體也會分清自己種族與其他種族之間的界限，在約會方面尤其如此。這令他非常驚訝。更進一步說，一個種族的文化基礎是根深蒂固的，需要數十年乃至更久的時間才

能得到調整，而約會只是在這種基礎上開展的一項活動，必然受到文化基礎的制約。

長期以來，很多白人從事所謂的種族科學研究，試圖證明白人優越論，但這種所謂的科學是一種令人尷尬的偽科學。我也非常清楚地知道諸如「女性發現白人男性更有吸引力」之類的數據是如何得來的。我不會說白人比其他種族的人好看得多，也沒有說我的數據表明黑人沒有吸引力。事實上，在美國以外的地方，OkCupid網站的模式存在一些變化。在英國，該網站的黑人用戶獲得的信息量是白人用戶的98.9%，這個數字在日本是97.8%，在加拿大是90%。在前兩個國家，尤其是日本，許多黑人用戶都是身在國外的美國人。

有時候，與性有關的問題與骨骼結構和肌肉無關。每個種族都有優點，也有不足。文化背景、期望值和適應能力都會產生一定的影響，這就是我們的數據揭示的結論，因為數據源自人與人之間，而且非常詳細，所以能夠以其他研究無法企及的方式幫助我們看清真實的情況。

在高中時期，有一年夏天，我以交換生的身份去日本，居住在本州島中部的宇都宮市。日本方面的組織者偶爾會帶領我和其他美國人去參觀附近的學校或工廠，以便讓我們儘可能多地看看這個國家，也讓這個國家的人們看看我們，增進雙方的理解。當時是20世紀90年代初期，互聯網時代還沒有到來，美國主要的經濟對手還是日本，而不是中國。美國和日本之間的關係有些緊張，在我赴日的幾年之前，日本買下了洛克菲勒中心，日元對美元構成了嚴重威脅。那次交流項目的名稱雖然只有三個單詞，卻體現了這次交流活動的宗旨，這三個單詞是「Youth for Understanding」（增進青年理解）。

儘管這個名字表明此次交流是促進理解的，但兩國的文化差異令我困惑不已。我清楚地記得，即便《街頭霸王II》這款遊戲中的一些人物的名稱在日本和美國都不同，日版中的維加（Vega）到了美版中被叫作巴爾羅格（Balrog），巴爾羅格被叫作比鬆（M.Bison）……這簡直瘋了。當時，日本也有美國電視劇，不久，《海灘救護隊》

（*Baywatch*）^[15]就成了日本獨佔鰲頭的電視劇。在一所學校裡，我們要站起來，當著臺下所有學生的面講幾句話。我走到講臺上傻傻地講了幾句之後就走下了講臺。接下來要在講臺上發言的學生是我們交換生中唯一一位金髮女郎。她站起來時，下面的學生中爆發出了一陣驚歎，驚歎聲甚至大到了我們能聽到的地步，我永遠不會忘記當時的情景。她不過是一個普通的女孩，我們都是16歲，顯得有些笨拙和膽怯，而人們看到

這個女孩時卻爆發出一陣驚歎，彷彿站在講臺上的那個人是個當紅女星。

許多人對這種驚歎的理解都只是停留在表面上。數十年來，顱相學家、種族主義者以及醫生都跳了出來，試圖從生物學角度來解釋其他種族的人看到白人時的驚歎，試圖將白人優越論牢固地確立下來。普林斯頓大學的美國史黑人女教授內爾·歐文·佩因特（Nell Irvin Painter）於2010年出版的《白人的歷史》對所謂的「種族科學」進行了非常好的概括。在這本書中，她引用了一段啟蒙運動時期白人得到的評價（不用說，肯定是白人寫的）：

格魯吉亞的血統或許是世界上最好的。我在那個國家沒有發現過一張醜陋的臉龐，無論是男性還是女性，我只見過天使般的人。大自然慷慨地將其他地方不可能看到的美貌賜予了那裡的女性.....不可能找到比格魯吉亞人更迷人的面孔或更好的身材了。

事實上，這段話的作者的確是一位白人，他是德國人類學家約翰·布盧門巴赫（Johann Blumenbach）。他通過收集和比較不同種族的顱骨形成了自己的種族理論。現在，種族研究可能已經進步了很多，而人們的潛意識卻是另外一回事。

[1] See, for example, 「Blacks Still Dying More from Cancer Than Whites,」 by Jordan Lite, Scientific American, February 2009. Also see the Sentencing Project's 「Criminal Justice Primer for the 111th Congress,」 which details many depressing disparities in the sentences handed down to whites, compared to minority defendants: sentencingproject.org/doc/publications/cjprimer2009.pdf.

[2] The headline cited is from ThinkProgress.org. 「Study: Black Defendants Are at Least 30% More Likely to Be Imprisoned Than White Defendants for the Same Crime,」 by Inimai Chettiar, August 30, 2012, thinkprogress.org/justice/2012/08/30/770501/studyblack-defendants-are-at-least-30-more-likely-to-be-imprisoned-than-white-defendants-for-the-same-crime.

[3] See Jason Lisk, 「Quarterbacks and Whether Race Matters,」 The Big Lead, December 2, 2010, thebiglead.com/2010/12/02/quarterbacks-and-whether-race-matters/. 當然，儘管我只發現這一位作者統計了不同種族四分衛獲得的評分，但這並不能證明其他人沒有做過相關統計，只不過我花了好幾個小時梳理搜索結果，只找到了這一篇文章而已。

[4] 當然，並不是OkCupid網站的每一個用戶都會把自己簡單地劃入這四大種族中，但為了便於討論，我將分析範圍侷限在了這四大種族。

[5] 在OkCupid網站上，15%的用戶選擇了不止一個種族，3%的用戶選擇的種族不在這四大種族之中。像這些選擇多個種族、選擇四大種族之外以及不選擇任何種族的用戶都被排除在了分析範圍之外。

[6] 我對每一欄的簡單平均數進行標準化，而不是加權平均數，因為白人佔了多數。如果將加權平均數標準化，就會導致扭曲，造成以白人為典範的假象。但如果將簡單平均數加以標準化，那麼捕捉到的意義就是這樣的：「A種族的甲遇上B種族的乙時，甲對乙的評分，與甲對其他種族的評分有何不同？」這是一個有趣的問題，我們打算做進一步研究。

[7] 黑人女性獲得私信的數量大概只有其他種族女性的75%，私信得到的回覆率大概也是75%。

[8] 在我們的樣本中，與非黑人男性對於其他種族女性的評分相比，非黑人男性對黑人女性的大多數評分都減少了0.2到1個星，對其偏差進行分析的結果表明，評分的中位數減少了0.6個星。82%的樣本都表明，歧視黑人的現象一直穩定地存在著。

[9] 雖然OkCupid的數據來自我們的內部資料，但如果想看到這些數據與Quantcast全國平均數的比較，請訪問<https://www.quantcast.com/okcupid.com?country=US>，然後點擊Demo graphics按鈕，從其菜單中選擇Ethnicity選項，然後打開US Average選項即可。

[10] 當然，交友網站並不是完美的數據庫。我們都知道，這類網站的用戶都是單身人士，會影響到一些調查結果。比如，如果要藉助這類網站去研究美國人的消費習慣，然後得出結論說美國人一般都會將可支配收入花到吃飯和看電影上，那就太偏離現實了。像這類沒有考慮到數據庫特徵的錯誤觀點，就會顯得荒誕不經。

[11] 下面這些故事都是由用戶自己上傳到OkCupid網站的Success Stories欄目。關於貝拉和帕特里克的故事，請訪問<https://www.okcupid.com/success/story?id=2855>；關於丹和簡恩的故事，請訪問：<https://www.okcupid.com/success/story?id=2587>。

[12] 這段話摘自奧巴馬對於喬治·齊默爾曼判決的評論，鏈接為：whitehouse.gov/the-press-office/2013/07/19/remarks-president-trayvon-martin。

[13] See 「Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination,」 by Mari-anne Bertrand and SendhilMullainathan, American Economic Review 94, no.4 (2004): 991—1013, doi:10.1257/0002828042002561. My discussion of Obasogie's work relies on Francie Latour's Boston Globe article 「How Blind People See Race,」 January 19, 2014. Latour provides a précis of Obasogie's book *Blinded by Sight: Seeing Race Through the Eyes of the Blind* (Redwood City, CA: Stanford University Press, 2014), and interviews him.

[14] 我在這裡需要澄清一下，我使用「我們」一詞，絕對不是矯情，而是說我自己的的確確也會存在這個問題。

[15] 1992年，我在日本，雖然當時這部電視劇已經在全球流行開了，但直到一年之後才登上日本的主流熒幕。然而，衝浪文化、加州和有著古銅色皮膚的金髮碧眼女郎等時尚元素早已風靡日本。如果你走進一家令人感覺挺「酷」的服飾店，可能會發現裡面正在播放海灘男孩（Beach Boys）的歌曲。記住，我講的是1992年的情景，海灘男孩的歌曲是「Surfin' Safari」，不是1988年的「Kokomo」。

第七章 被神化的美貌

在我工作的領域裡，人們會根據每一個能想到的標準來劃分自己和其他人，比如，抽菸者和不抽菸者，基督徒和無神論者，傻子與天才，黑人、白人與亞裔，男同、女同、雙性戀和性取向正常者等。人類就是由一個又一個部落組成的。或者用更加美麗的詞語來說，就是朝鮮諺語所說的那樣：「千山之外還是山。」^[1]這句話道出了朝鮮半島崎嶇險峻的地理特徵，也體現了生存地域的斷裂給人類造成的無盡艱辛。

在經營交友網站時，你會發現人與人之間存在這樣一個細分標準，這個標準似乎沒有依據，又像一個人的種族或性別一樣是與生俱來的，而且往往又不容易對其進行直接分析。在OkCupid網站上，如同Match.com和Tinder等約會網站一樣，用戶之間一個主要的，或許也是最深刻的區分標準就是外貌之美。有些人擁有美貌，而有的人沒有。在性吸引力上，從父母那裡繼承了美貌的人佔盡優勢，而其貌不揚者卻基本上什麼好處也沒有。美貌與種族一樣，也是你手中的一張牌，會產生巨大的影響。

下面，我根據用戶的外貌吸引力繪製了一幅圖，反映了用戶每週新收到的信息量（見圖7—1）。

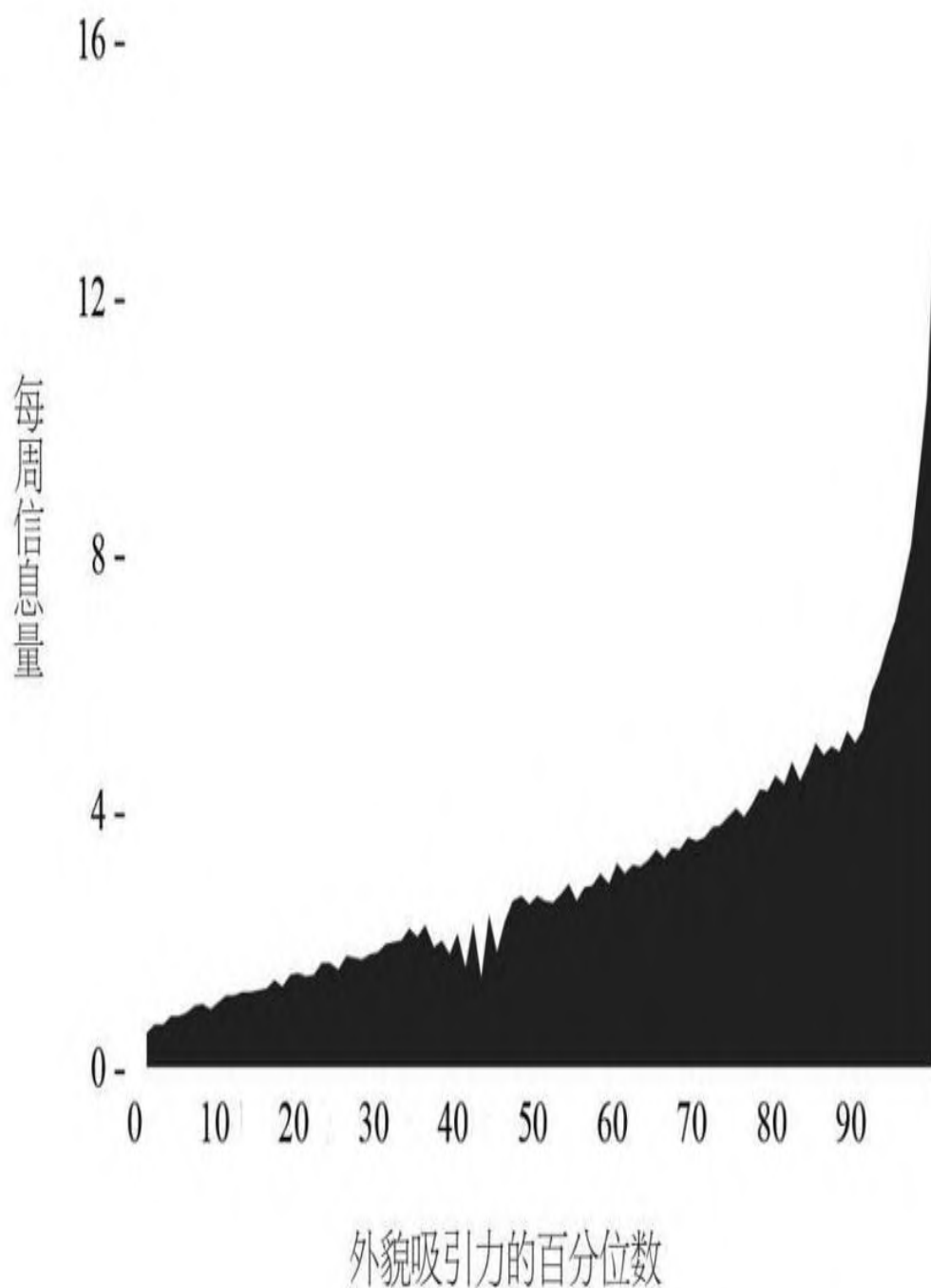


圖7—1 用戶每週收到的信息量

圖7—1中右側曲線的驟然上升態勢遠遠壓倒了其他部分，因此，它的真實性在一定程度上被掩飾。從最低的百分位往上看，大致可以用一個指數函數來描述。也就是說，地震學家用來測算地震釋放的能量時所用的算法也適用於這個圖形，我們可以用里氏震級來計算美貌產生的影響。^[2]如果震級為1.0或2.0，那麼其產生的影響幾乎沒有什麼明顯區

別，震動無法為人感知到。但如果震級非常高，一個小小的差異就會產生災難性的影響。如果震級為9.0，就會產生強烈震動；如果為10.0，就會震碎整個世界。

由於我在圖7—1中將男性與女性獲得的信息量彙總在了一起，掩蓋了兩性之間的差異，所以，你肯定不會發現這樣一個結論，即男性與女性對美貌的體驗是不均衡的。在圖7—2中，我分開呈現男性與女性每週新獲得的信息量，中間那條虛線表示彙總在一起後的變化態勢。

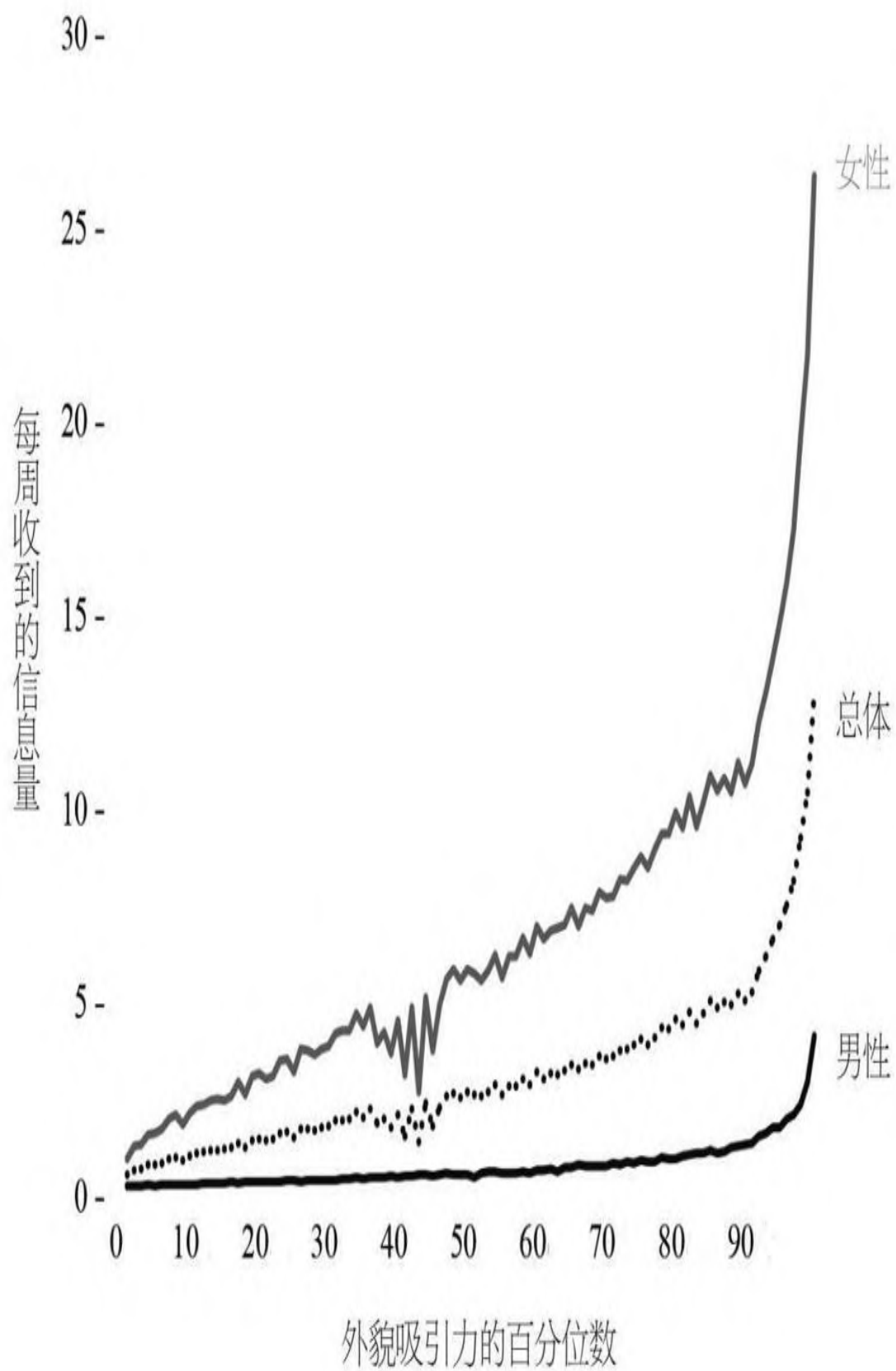


圖7—2 不同性別每週收到的信息量

在圖7—2中這條灰色曲線的右上角，隨著外貌吸引力的增加，我很難準確地描述出這些女性用戶每週收到了多少條信息。她們根本無須追著你，大聲告訴你她們的愛好。大城市裡的信息流比我們在上面看到的要多出50%，對於那些外貌吸引力最強的女性而言，每天都有很多男性在等著她們上線。這些男性用戶會給她們發來一些類似於「嗨，我喜歡摩托車，你喜歡嗎？」之類的搭訕信息，甚至可以說，等待這些女性的男士如滔滔江水般綿延不絕。然而，美貌的影響以及兩性的差異不僅僅體現在與性有關的領域。

圖7—3的數據來自Shiftgig（一家面向小時工和服務業的招聘網站），我研究的樣本數量大概是5 000人。^[3]這些數據揭示了男性與女性外貌吸引力對於其面試機會的影響。^[4]

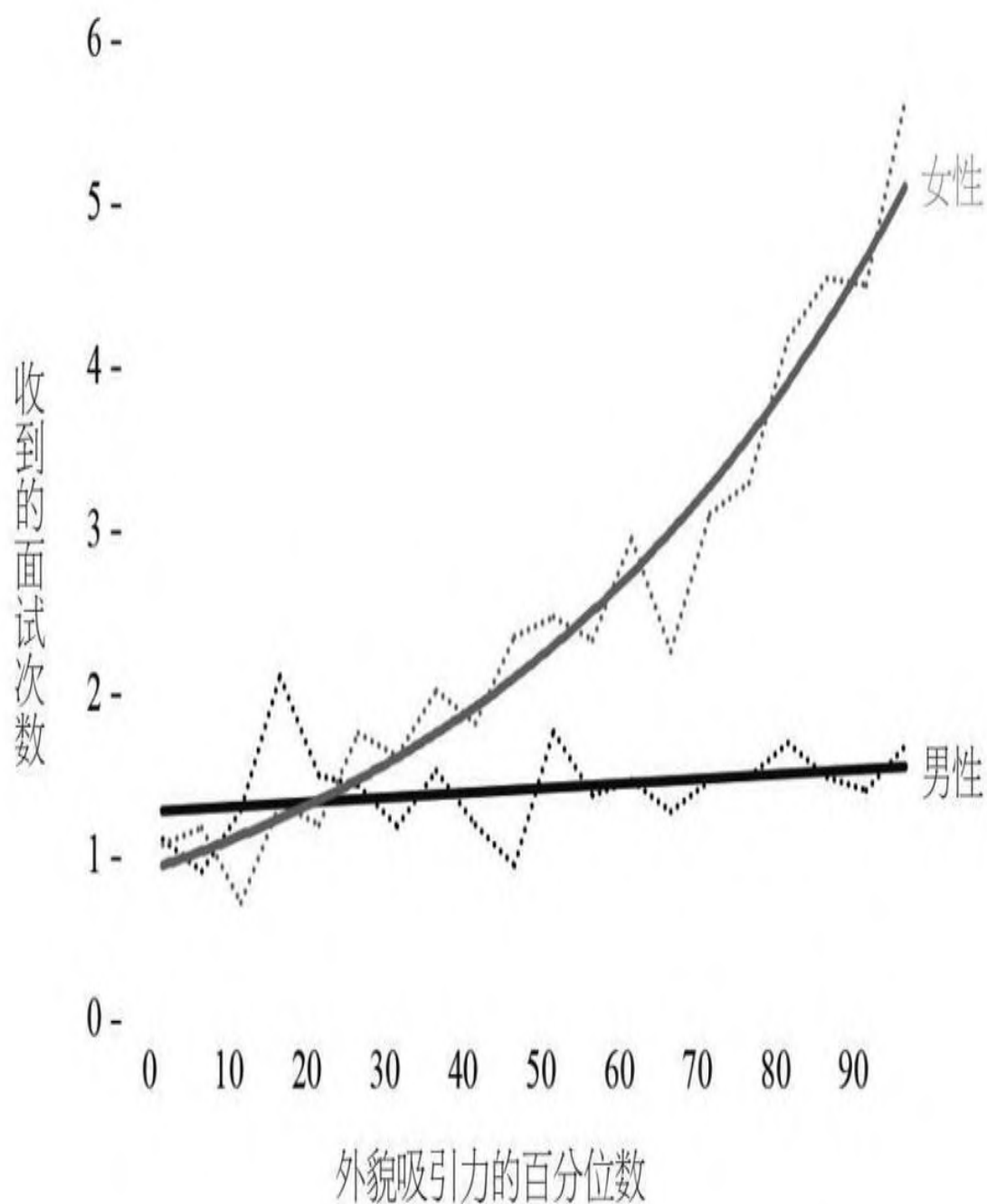


圖7—3 外貌對面試機會的影響

外貌吸引力也會影響到男性與女性在Facebook上朋友圈的大小，如圖7—4所示：[\[5\]](#)

無論對於男性而言還是對於女性而言，成功與外貌都存在關聯性，但正如你看到的那樣，表示女性的灰色曲線的坡度最大。在Facebook

上，男性的外貌吸引力每增加一個百分位，他的新朋友就會相應地增加兩位；而女性的外貌吸引力每增加一個百分位，則其新朋友就會相應地增加三位。在Shiftgig網站上，兩條曲線已經無法通過這種方式來對比了。女性的曲線是指數型的，而男性的曲線則是線性的。此外，無論招聘經理是男性還是女性，這兩條曲線都是適用的。男性應聘者那條曲線幾乎是水平的，這說明男性的外貌對其面試機會幾乎沒有什麼影響；而女性那條曲線則是指數型的，這就表明雖然女性是在投簡歷、找工作，但她們仍然能夠像在OkCupid網站上那樣受到青睞。男性招聘者在權衡女性應聘者時，會考慮到其外貌吸引力，就像在一種浪漫關係中考慮女方外貌一樣：要麼令人沮喪，要麼令人激動，這取決於女性應聘者的外貌和職業。如果女性應聘者曾經是一名律師，而且處理過訴訟業務，那麼可能影響男性招聘者對她的感受。女性招聘者似乎也會通過性的角度來審視男性應聘者，但一般情況下，不會帶著浪漫環境下的那種意圖。

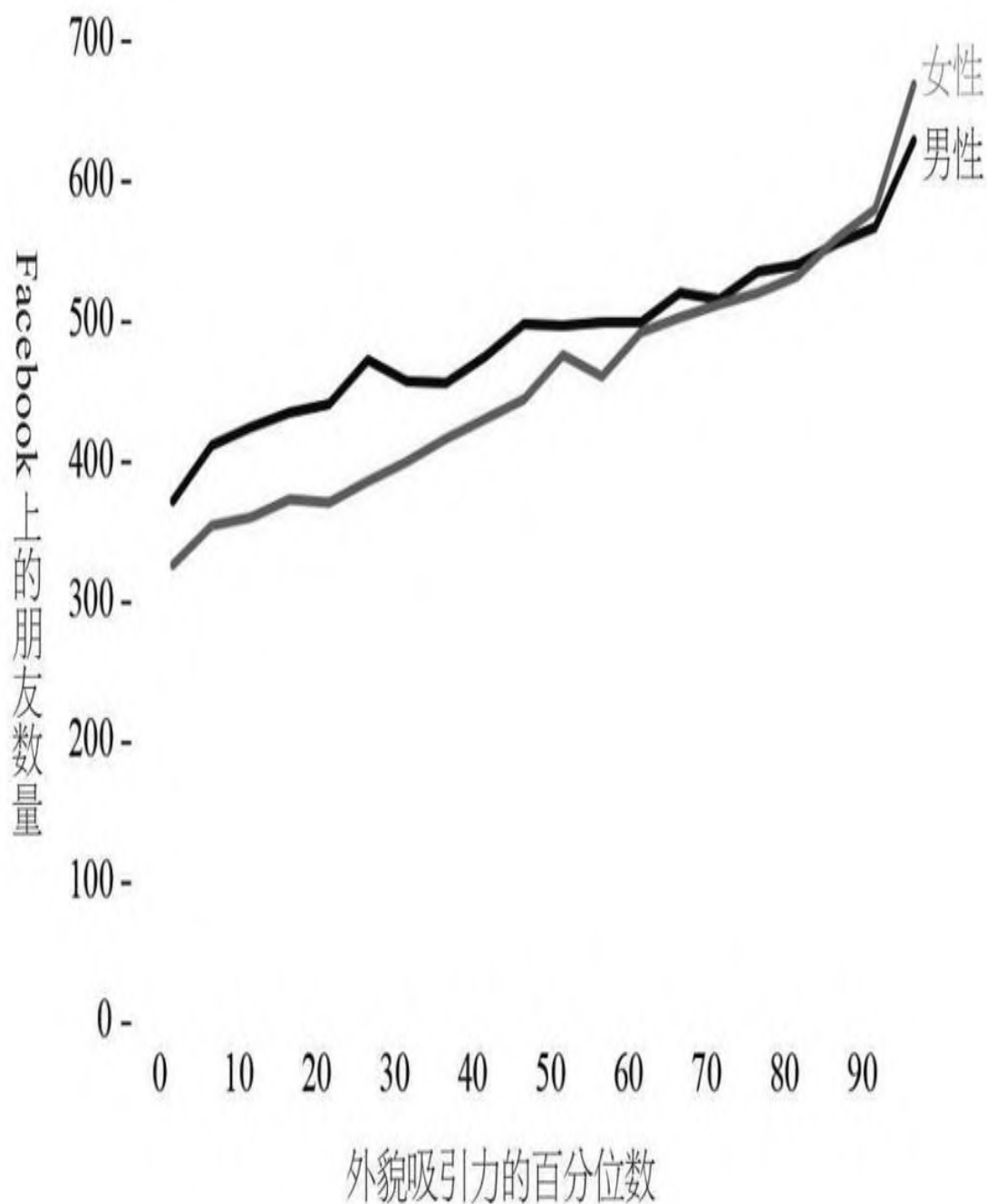


圖7—4 外貌對Facebook朋友圈的影響

的確，美貌很重要，對女性更是如此，而且這個現象由來已久。比如，社會心理學領域的一篇具有基礎意義的論文的題目就是《美的就是好的》。^[6]這篇論文發表於1972年，它以及一系列後續研究都印證了美貌的重要性。^[7]擁有美貌的人被視為比其他人更聰明、更有能力、更值

得信賴。更有吸引力的人容易獲得更好的工作。在法庭上，外貌較好者更容易被無罪釋放或者獲得輕判。羅伯特·薩波爾斯基（Robert Sapolsky）在《華爾街日報》上撰文指出，杜克大學的兩位神經心理學家正在研究為什麼大腦的眼窩前額皮質同時參與對臉部美麗程度和行為優秀程度的評價，而且在開展其中一項任務時，這個腦區的活動水平預測了開展另一活動時的活動水平。換句話講，大腦認為一個人的顴骨在一定程度上反映了其思想和心靈。從神經學角度來看，人類大腦對性魅力的記憶能力是非常強大的，一個性感的女性會立即給大腦留下深刻的印象。

美貌對女性影響尤其大。納奧米·沃爾夫（Naomi Wolf）的《美貌的神話》（*The Beauty Myth*）一書對這一點進行了更好的說明。因此，我所發現的這些現象並不新鮮，但與之前相比，我們現在卻能借助數百萬人的行為來驗證固有觀念，甚至一些人所共識的觀念。這種基於大數據的研究能為之前的工作提供一些輔助，有助於我們從中發現細微差別，甚至能幫助我們在之前工作的基礎上開展更加深入的研究。

《美的就是好的》這篇論文的作者只調查了60個人，遠遠無法為其結論提供充分的佐證，更無法證明美貌在多個領域的影響。但現在有了大數據之後，我們不僅可以充分證明美的就是好的，還能證明美貌在多個方面究竟有多好。比如，在性方面，美貌是非常好的；在友誼方面，美貌只是有點好；在找工作方面，美貌的效果則取決於性別。對於沃爾夫所做的開創性工作，我們能利用大數據證明她在廣泛觀察的基礎上提出的一個結論，即「子宮禁錮了維多利亞時代的女性，而美貌禁錮了今天的女性」，我們之前從三個交友網站上引用的數據就印證了美貌程度與女性所獲青睞之間具有密切的關係。此外，我們還能更加深入地證明她提出的關於「美貌是一種社會控制手段」的觀點。我們可以想一下來自招聘網站Shiftgig的數據，這些數據改變了我們對職場女性表現的理解。在很多情況下，她們之所以被遴選出來，並不是因為她們有能力做好一份工作，而是因為一個與工作能力基本沒有關聯的特質——美貌。與此同時，男性則沒有這樣的機會。

因此，從整體來看，與男性相比，女性在職場上註定更加容易失敗，這是一個簡單概率事件。之所以出現這種現象，一個至關重要的原因就在於遴選標準，而不是女性自身。想象一下，如果在招聘過程中，不論男性的工作能力如何，僅僅把體力作為遴選標準，那麼最終必然出現這樣的情況：男性雖然力氣大，卻面臨著一些無法用蠻力來應對的挑戰。同樣的道理，在僱用女性的過程中以貌取人就決定了她們在職場上

表現不佳（至少從統計學上來看是這樣）。可以說，女性施展才華的機會之所以受到了限制，原因就在於遴選標準和負責遴選工作的男性。因此，沃爾夫說：「事實上，美貌而不是外貌，總是在規定著行為。」她這句話主要是在與性有關的環境中說的，但我們現在用數學方法分析了美貌在職場環境中的作用。

我在前面提到過我有一個女兒。我和妻子萊西瑪難得同時放下工作，不過我們一有時間就會坐下來聊天，展望一下女兒未來的生活。如果有片刻清靜，所有父母難免都會這麼做，就像兩個醉漢在酒吧裡總是難免會爭辯。每個家庭都會這樣想象，想象的內容不盡相同，不過我想我們家的內容與大多數家庭的內容是一樣的。我和妻子會說（誰第一個說並不重要）：我們的小女孩那麼聰明。哦，是的，我們要儘可能把一切都教給她。她會變得非常溫柔、非常有愛心，這些對美好生活非常重要，我們在這些方面的看法是一致的。肯定會的，看看她的皮膚，像奶茶一樣，看看她的眼睛，我想她以後肯定會特別漂亮，因此，在她十來歲的時候，我們要儘量限制她外出，把她保護好。對於這一點，我和妻子出現了一些分歧。我妻子覺得女兒未必會變得像我想的那麼漂亮，因此，我們不會限制女兒的活動。我們坐下來，然後就開始聊其他話題。聊完天我想，如果是個兒子的話，我無法想象父母想給他施加限制。

不過，互聯網無疑讓我擔心的問題變得更糟。社交媒體上充斥著有關美貌的海量圖片，似乎預示著美貌的神話走向了終結。在每一份簡歷、每個應用程序、每一篇署名文章上，幾乎都能找到照片。如果有人關心你在做什麼，他們就能在網上找到你的照片。Facebook和領英這兩個社交網站無處不在的照片，不僅使愛情變得盲目，其他一切也變得盲目了，因為美麗的外貌幾乎壓倒了一切。在10年前，想通過一個普通人的名字去搜索他的照片是幾乎不可能的事；但現在，你只要在谷歌中輸入一個人的名字，瞬間就會搜索出他在某個社交網站上的照片。我們不得不根據這些快照選擇「最漂亮」的那一位。不過，在選擇過程中一定要明智，因為現在照片的展示方式與以往不同。社交網站內有個趨勢，行外人可能沒有清楚地意識到。近兩三年來，新的網頁設計標準變得更開放，照片所佔的空間更大，位置更突出。這樣一來，不僅照片變得更重要，而且美貌也更重要了。OkCupid網站最近就對照片顯示方式做出了一個調整，即照片所佔的位置從圖7—5中的黑框擴大到了紅框。

設計師們只是想讓頁面看起來更現代。他們沒有預料到這個結果：多餘的像素使得漂亮的頭像更炫目，而其他人的頭像則相形見绌，以至出現了「強者愈強、弱者愈弱」的馬太效應。在這方面，這種網頁設計

理念與美國的國內政策具有一定的相似之處。後來，設計師們不得不想辦法緩解這個問題（見圖7—6）。

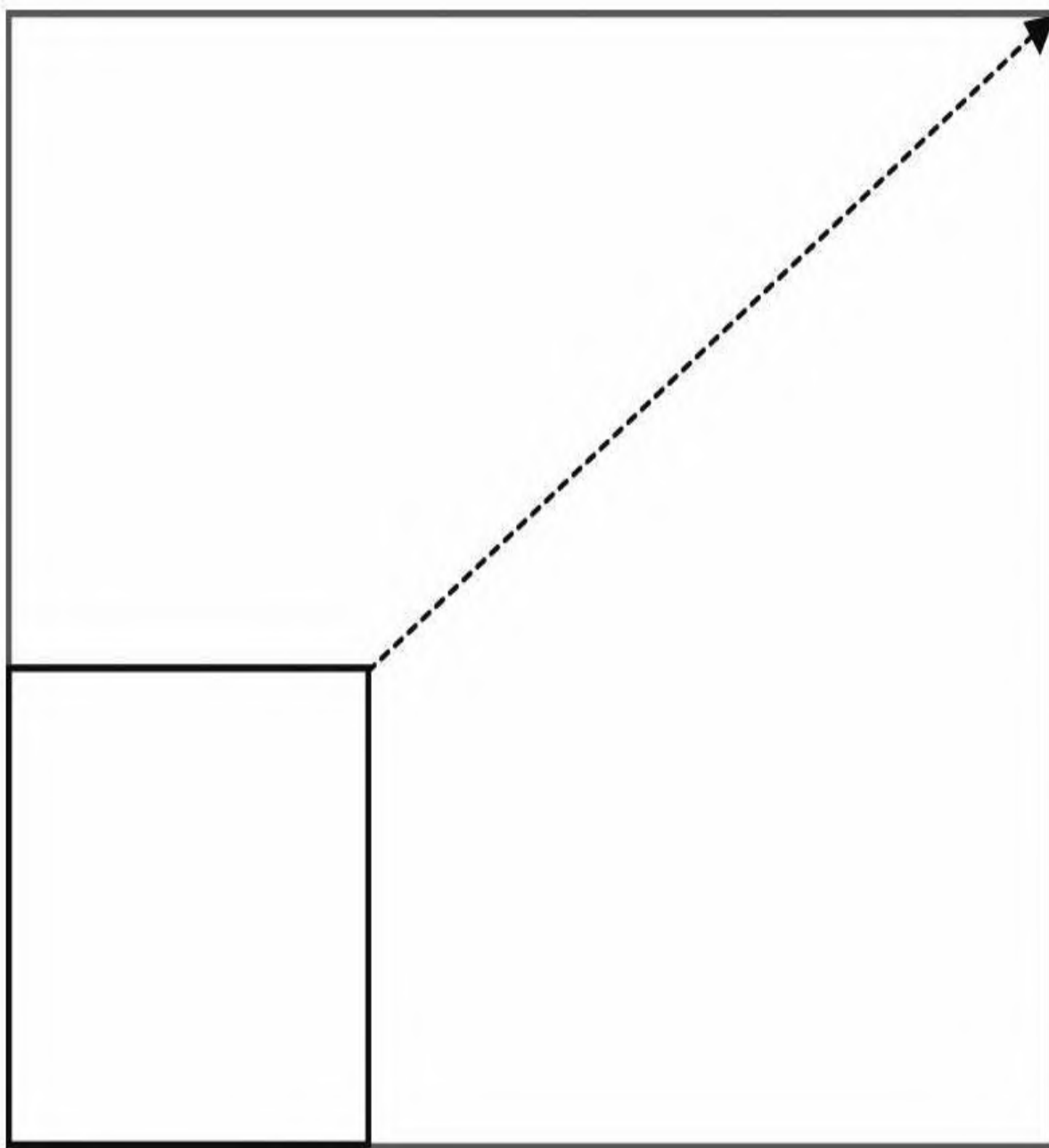


圖7—5 OkCupid網站的照片調整

考慮到這種壓力，發佈照片的博客大行其道便不足為奇了。帶有「勵瘦」（thinspiration）、「減肥」（lose weight）、「持續減肥」（keep losing）、「節食運動」（proana）、「大腿間距」（thigh gap）

等標籤的圖片充斥社交網站，以致Tumblr和Pinterest等網站不得不先後修改服務條款，禁止用戶上傳這類內容。^[8]如果你不知道最後兩個標籤的含義，那麼「節食運動」意為「支持通過節食來減肥」，而「大腿間距」的意思是由於你兩條腿非常細，當你雙腿站直，左右腳和膝蓋都併攏時，兩條大腿之間彼此碰觸不到，這被認為是好身材的標誌，也是青春期的少女盲目追求的一個標準。這種願望固然是美好的，但從生理學角度來看，大多數人都是不能如願以償的。如果利用谷歌搜索一下上面這些關鍵詞，你就會面對無數張在鏡頭前搔首弄姿的女性的照片，有的照片只顯示了身體的某個部位，那些「勵瘦」女性不僅非常瘦，往往還穿著睡衣、比基尼或內衣。這些女性創建的博客其實反映了男性的關注。我之所以會這麼說，是因為我對學界左派的語言本能地有所質疑。

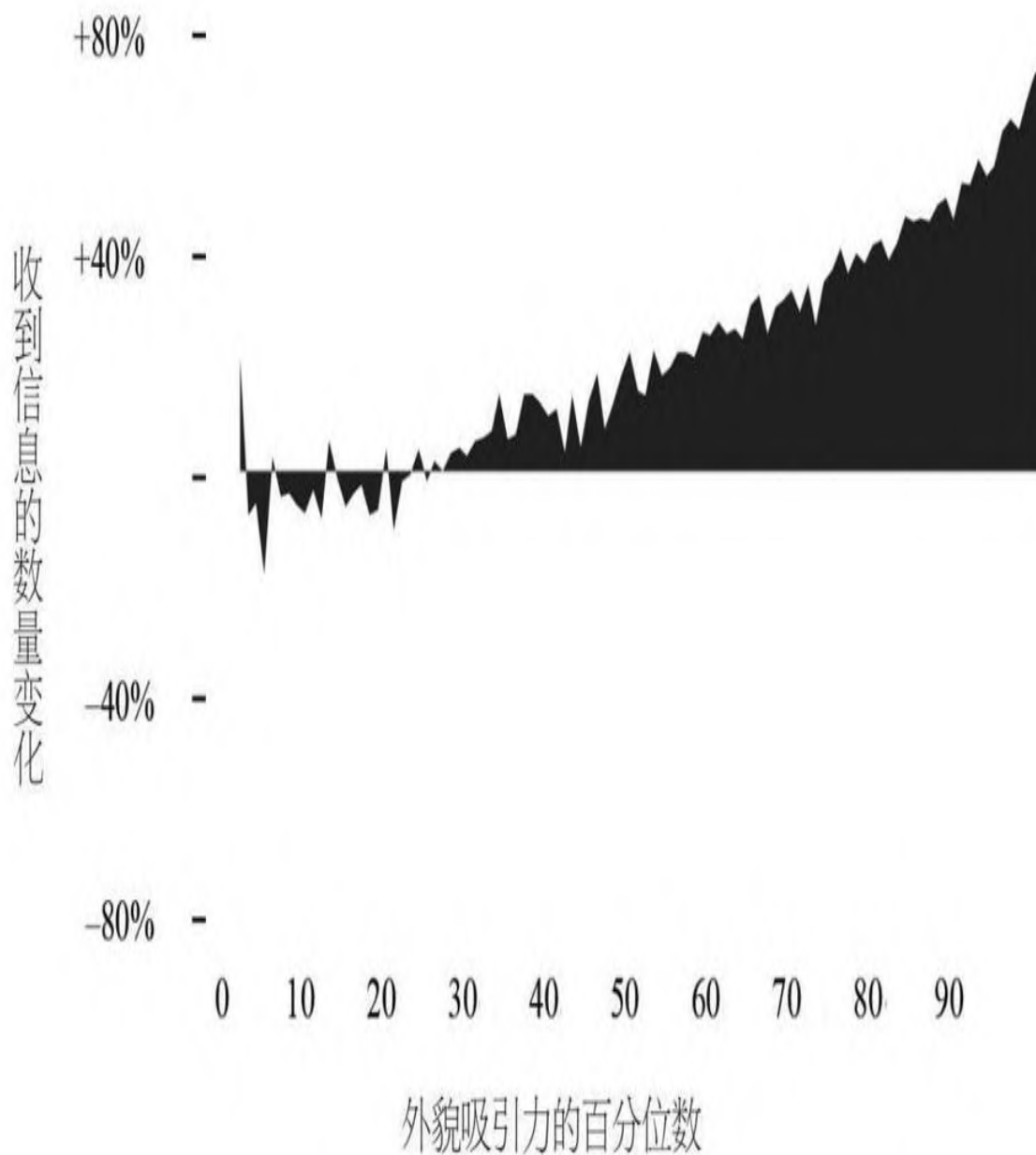


圖7—6 照片像素調整後用戶收到信息的數量變化

當然，雖然Tumblr和Pinterest針對這類內容制定了禁令，但問題並沒有得到解決，最起碼它們的用戶仍然在上傳一些性感裸露的照片。於是，這些網站現在便採取了另外一種方法。因為那些博文都有標籤，所以網站管理者能夠運用數學算法加以干預。比如，如果你在這些網站上搜索「大腿間距」一詞，那麼電腦屏幕上就會空白，上面就會出現這樣一行字：「如果你或你認識的某人患有飲食障礙症……」這行字的下面會給出一個鏈接，點擊這個鏈接，就能獲取一些幫助你克服飲食障礙症

的建議和資源。這只是人們在反對過度減肥的問題時採取的一個小措施，但在人們過度追求瘦身的行為並以數字化的方式表現出來之前，幾乎沒有人直接採取過任何措施來解決這個問題，至少在其危害顯現之前沒有。人們只是聽到一些傳言，也許有些父母也會對這種做法提出悲傷的質疑。數據展現的是我們的真實感受，包括我們對他人以及對自己的感受。如果文化、政治、習慣、部落等方面的數據存在差異，這就說明我們人類自身也存在差異。這種想法給我們帶來了希望，因為無論對於任何事物，如果想要使其完整，第一步就是要知道缺失了什麼。

[1] 這句諺語摘自威廉·曼徹斯特為麥克阿瑟將軍撰寫的回憶錄《美國愷撒》（American Caesar）（New York: Little, Brown, 1978）。我之所以會讀這本回憶錄，是因為我在寫作過程非常艱難的時候，想讀一點無關的資料，將我的注意力從繁雜的數據上轉移開來。

[2] 我當時已經非常熟悉里氏震級的對數本質，但還是在維基百科上查閱了Richter magnitude scale（里氏震級）的條目，以便了解標準的震級究竟有什麼內涵。當然，把美貌與震級相比，有點像詩人那樣採取了一種誇張的描述手法，畢竟二者產生影響的方式並不完全一樣。

[3] 相關數據由該網站的數據團隊提供，該網站創始人埃迪·盧（Eddie Lou）非常積極地配合我。

[4] 我在這裡加入了趨勢線的原因是這裡的樣本大約是5 000人，與一般的圖相比，這張圖顯得較為分散稀疏，不易觀察總體趨勢。

[5] OkCupid網站的部分用戶選擇將其賬戶與Facebook賬戶捆綁在一起，這是我們對有關數據加以整合與匿名化處理之後得到的結論。

[6] See 「What Is Beautiful Is Good,」 by Karen Dion, Ellen Berscheid, and Elaine Walster in *Journal of Personality and Social Psychology* 24 (1972): 285—90.

[7] This passage adapts conclusions from and directly quotes 「Pretty Smart? Why We Equate Beauty with Truth,」 by Robert M. Sapolsky, in the *Wall Street Journal*, January 17, 2014. The Duke neuropsychologists alluded to are Takashi Tsukiura and Roberto Cabeza. See also 「Jurors Biased in Sentencing Decisions by the Attractiveness of the Defendant」 at *Psychology and Crime News* for an overview of the effects of physical attractiveness in the criminal justice process: [crimepsychblog .com/?p=1437](http://crimepsychblog.com/?p=1437), posted by user EmmaB, April 3, 2007.

[8] See 「A New Policy Against Self Harm Blogs,」 Tumblr's staff blog, March 1, 2012, staff.tumblr.com/post/18132624829/self-harm-blogs. See also 「Pinterest Thinspiration Content Banned According to New Acceptable Use Policy,」 by Ellie Krupnick, *Huffington Post*, March 26, 2012, huffingtonpost.com/2012/03/26/pinterest-thinspiration-content-banned_n_1380484.html. The *Huffington Post* has actively covered the 「thinspiration」 phenomenon. See 「Th H Bl A St Wld f T eungerogs: ecreooreenage「Thinspiration,」」 by Carolyn Gregoire, February 8, 2012, huffingtonpost.com/2012/02/08/thinspiration-blogs_n1264459.html. For more on 「thigh gap」 (and for evidence that altering the Terms of Service did not solve the problem), see 「The Sexualization of the Thigh Gap,」 by Allie Jones, on *The Wire*, November 22, 2013, thewire.com/culture/2013/11/sexualization-thigh-gap/355434/.

第八章 隱祕的選擇

長期以來，若要知道一個人的真實想法，通常有兩種途徑。一種是在對方沒有防備的時刻去窺探他，如果你們在實驗室裡，也可以用實驗服做掩護，或用其他辦法讓對方忘掉你正在觀察他。像這類事情可能充滿樂趣，你可能會用到實驗服、隱藏的攝像頭，也可能用上假鬍子等道具。但總體來看，這種方法是不大可能行得通的。因此，如果你想獲得數據，一般情況下需要選擇第二種途徑：向對方提出問題，希望得到真誠的回答。自從喬治·蓋洛普（George Gallup）於1935年創辦美國輿論研究所以來，第二種途徑一直是比較流行的調查方法。^[1]

不過，從歷史上看，在涉及種族、性行為、吸毒甚至身體功能等話題時，調查結果往往無法反映出人們的真實態度，因為受訪者會對他們的回答進行加工和編輯。^[2]觀察到的行為數據是非常有用的，我們在前文已經看到了這一點。但有些事情，比如思想、信念等，不需要通過一個明確的行動來表現，而分歧最大的、最醜陋的態度往往隱藏在自我意識和文化規範的面紗後面，而在開展調查的過程中，至少在直接詢問的調查過程中，受訪者很難擺脫這些面紗來展現真實的自我。你最想獲取的信息恰恰是受訪者最想隱藏起來的，這是社會科學家們面臨的一個詛咒。這種傾向被稱為「社會期望偏差」。這方面的案例非常多，在世界各地都存在這種現象，受訪者總是會注意自己回答問題的方式，以便給別人留下好印象。^[3]最著名的一個例證就是所謂的「布萊德利效應」

（Bradley effect）^[4]。1982年，作為民主黨的黑人湯姆·布萊德利（Tom Bradley）競選加利福尼亞州州長。在選舉結束後的民意調查中，該州的很多選民都對調查者說自己把票投給了布萊德利，以至布萊德利的支持率大幅領先於其對手——一位白人競選者，但其實他們悄悄地把票投給了白人競選者，最終使那位白人以微弱優勢勝出。在整個20世紀80年代和90年代，黑人競選者在民意調查中得到的支持經常高於其在實際投票中得到的支持。之所以出現這種現象，就是因為白人選民不願意表現出自己種族歧視的一面，通常會向調查者撒謊，表示會支持黑人競選者，但在實際投票時，他們仍然會選擇白人。除了種族問題以外，諸如抑鬱症、上癮症等問題也難以從社會層面去加以評估，因為人們不會誠實地

回答。即便在OkCupid網站上，也存在這種現象。一般來講，網站為了幫助用戶尋找匹配對象而提出一些問題時，除了回答者本人以外，其他人不會看到他回答了什麼。即便如此，用戶也不願意透露出真實態度，但在這個網站的其他環節，用戶的行為會洩露出其真實態度。你一提出問題，用戶在回答之前就會主動審查一下自己的回答，然後給出經過加工的版本，而非真實的版本。幾乎每一個記錄用戶態度的網站或收集描述性數據的網站都存在同樣的問題。但有一個地方卻不需要問任何問題，因此其收集的數據是完全真實的：用戶主動搜索內容時，雖然沒有回答任何問題，卻在無意識的狀態下洩露了其真實想法。

谷歌就是這樣。打開谷歌網站後，其首頁上唯一一個具有提示性質的內容就是那個孤零零、空蕩蕩的長條形搜索框，光標已經放在搜索框裡了，只等你輸入要搜索的內容，一旦輸入，你就會暴露出自己的想法。谷歌公司的業務是幫助用戶從浩如煙海的互聯網信息中找出自己需要的信息，在這一點上，它的確做得非常成功。但在獲得舉世無雙的成功之後，谷歌卻記錄了人類的集體身份認同，因為它將人們輸入的每一條搜索內容儲存了起來。我們遇到問題時就會想到用谷歌搜索答案，谷歌儼然已經成為我們的醫生、牧師、心理醫生和密友。最重要的是，谷歌不會問我們任何問題，因為它那空白的搜索框已經隱藏了一個問題：嗨，你腦子裡想的是什麼呢？亞哈（Ahab）船長想的肯定是找白鯨復仇，亞瑟王想的肯定是聖盃。一個人搜索的內容往往會展現出真實的自己。那麼，現在的問題是：我們怎樣才能看到他人的搜索內容呢？

自2008年以來，谷歌公司推出了「谷歌趨勢」（Google Trends），通過對一段時間內的關鍵詞搜索次數及變化趨勢進行統計，總結出這段時間內的熱門內容。有了這個工具，任何人都能查詢之前只有網站管理者才能訪問的聚合搜索數據庫。輸入正確的關鍵詞，你就會看到與這個詞有關的列表，從而很好地窺探到廣大用戶內心的真實想法。其實，到目前為止，民意調查自始至終都無法獲取人們內心的真實想法。這個服務推出之後，科學家們用它來預測股市，探索經濟效率的驅動因素（他們發現富裕國家更關注未來，而非過去），^[5]實時追蹤流感、登革熱等疾病的蔓延情況，因此有利於迅速採取應對措施。^[6]當人們生病的時候，他們就會搜索有關的症狀和療法，這時，「谷歌流感趨勢系統」就會展開跟蹤分析，創建地區流感圖表和流感地圖，評估疾病蔓延情況，為美國疾病控制與預防中心提供預警。

谷歌還記錄了其他類型的「病毒」。因為谷歌不會提出任何問題，

而且不像社交網站那樣有人看到自己的信息，人們利用谷歌搜索引擎時會釋放出一些最不光彩的衝動。比如，**nigger**（黑鬼）就是一個常見的搜索詞，每年多達700萬次搜索包含這個單詞。^[7]在美國，這個詞搜索量最大的地方是西弗吉尼亞州，你或許能夠猜到這一點。但從整個美國來講，這個詞的搜索量一直都是居高不下的。紐約的布魯克林區與我成長的那個小城市——阿肯色州首府小石城幾乎沒有多少共同之處，但在這兩個地方，「**nigger**」這個單詞的搜索量卻是大抵相當的。此外，在伊利諾伊州的芝加哥市和加利福尼亞州的弗雷斯諾市，這個詞的搜索量也是大抵相當的。^[8]在美國，**nigger**比apple pie（蘋果派）的搜索量還要大，前者比後者多出近30%。^[9]此外，有確鑿數據表明，**nigger**這個對黑人的蔑稱在谷歌上出現的頻率，比在Twitter之類的社交網站上出現的頻率高出30倍左右。^[10]

疾病的蔓延情況具有顯著的週期性，而種族主義的暗流則從未停止過，這是一個長期存在的問題。疾病的蔓延是以人體的新陳代謝為基礎的，而種族主義的蔓延則依賴於一代又一代人。我們可以看到，與種族主義有關的數據在不同時期呈現出大幅波動。此外，如果我們把數據的跌宕起伏同現實世界中的事件聯繫在一起，就能夠挖掘出這些數據背後的人類情感。比如，如果你在2008年美國大選前後在谷歌上搜索**nigger**這個詞，並繪製出搜索頻率的變化圖，就能發現美國人民在是否選舉黑人為總統的問題上非常糾結。

從左至右，你會從圖8—1看到6個黑點，表示這個詞的搜索量飆升到了一個峰值。第一個峰值是超級星期二——2月5日。年初預選時，20多個州集中在這一天進行初選，其結果會對最終的黨內提名產生重要影響。接下來是4月22日賓夕法尼亞州的初選。希拉里在賓夕法尼亞州的民主黨初選中以9%的優勢擊敗對手奧巴馬。到了6月6日，搜索量達到了新高。希拉里宣佈退出競選，奧巴馬獲得民主黨總統候選人提名。7月15日，美國說唱歌手納斯（Nas）推出了一個名為《黑鬼》

（*Nigger*）的新專輯，導致**nigger**一詞的搜索量一度飆升，出現了第四個黑點所示的峰值。在整個分析過程中，突然出現這一事件，就構成了我們之前討論過的「混淆變量」。在這一事件之後，搜索量迅速減少，因為奧巴馬已經贏得了民主黨的總統候選人提名，種族以及政治話題引發的緊張態勢趨於緩解。事實上，9月初共和黨全國代表大會前後一週時間裡，與種族歧視有關的詞語的搜索量降到了整個選舉期間的最低點。^[11]

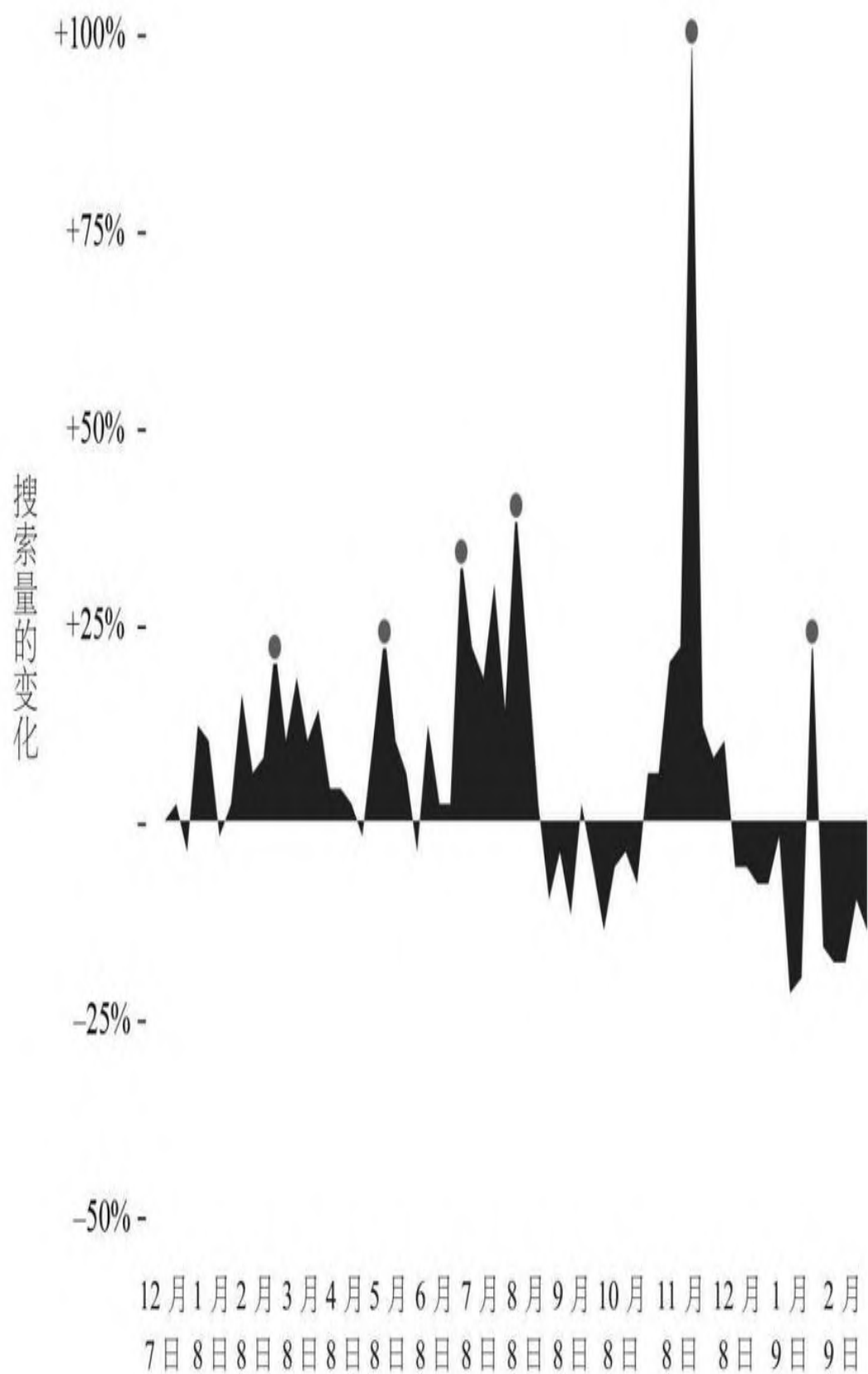


圖8—1 「Nigger」谷歌搜索量的變化（2007年12月—2009年2月）

然而，充滿敵意的搜索量在達到最低點之後，卻迅速反彈到了正常水平，之後在大選之夜迅速飆升到了一個空前絕後的峰值。第二天，當美國人一覺醒來，黑人成為總統的事實已經成為定局，帶有Obama字樣的搜索字符串中，大約有1%還包括nigger或KKK^[12]。在這個峰值之後，與種族歧視有關的搜索量幾乎立即出現了暴跌。^[13]到奧巴馬就職儀式時，針對黑人的憤怒出現了最後一次高漲，之後便一直保持在低於正常值25%的水平。^[14]美國一直在提倡推動不同種族之間的對話，削弱種族主義的影響，但看一看這些數據，你就發現其實對話並沒有多少效果，種族問題仍然根深蒂固。在這些數據中，你還能看到，雖然奧巴馬並沒有兌現其在競選期間一再承諾的變革，但他的就職的確改變了nigger一詞的搜索量。這個詞是美國人最愛搜索的單詞，但與奧巴馬就職之前相比，現在的搜索量已經大大減少了（見圖8—2）。

奧巴馬當選為總統之後，nigger一詞的搜索量出現過三次真正的攀升。第一次是2011年10月的第一週，共和黨總統初選候選人、得克薩斯州州長佩裡捲入了一場種族歧視風波。當時，其對手指責他與父親租用的狩獵營地有一個歧視黑人的名字，即「黑鬼頭」（Niggerhead），佩裡稱其父租下這個營地不久，已用油漆將字跡遮蓋，但有數名目擊者稱，數年後，字跡仍清晰可見。事件曝光之後，佩裡的對手、清談節目主持人凱恩和黑人牧師沙普頓等皆抨擊佩裡在種族問題上遲鈍、不敏感。後兩次攀升是由同一個事件引起的。其中，第二次攀升發生於2012年3月下旬，佛羅里達州黑人特雷沃恩·馬丁（Trayvon Martin）被白人喬治·齊默爾曼（George Zimmerman）槍殺一案引發了全美關注。第三次發生於2013年6月的最後一週，當時，當地檢察官拒絕起訴齊默爾曼。自從奧巴馬首次當選美國總統之後，白人可能在這兩個時刻感覺自己受到了莫大的威脅。在整個調查取證階段以及在2013年7月13日齊默爾曼被宣佈無罪時，nigger的搜索量都沒有出現類似的大幅攀升。如同2008年總統大選期間的情況一樣，在齊默爾曼無罪釋放之後，搜索量達到了新低，再次表明美國人在種族關係上的情緒存在週期性。

2009年1月20日之前

73

2009年1月20日之后

55

圖8—2 「Nigger」的谷歌搜索指數（按照日期）

如果你要搜索一些具有種族歧視色彩的詞，那麼顯然要從nigger一詞開始。但很快你就會發現，除了這個詞之外，真的很難找到內涵如此豐富的歧視性語言了。這個詞似乎是仇恨言論的首選。其他一些類似的蔑稱，比如「spic」（西班牙佬）和「chink」（中國佬）等，用得都非常少，可以拿來分析的數據也比較少。^[15]不過，最有意義的並非這些詞本身，而是它們背後隱藏的思維。比如，「nigger」一詞的內涵會隨著講話者的身份而變化。如果2008年發行《黑鬼》（*Nigger*）新專輯的不是納斯這個黑人，而是白人歌手託比·凱思（Toby Keith），那麼最後你看到的結局就截然不同了。在搜索這類詞時，谷歌自動完成功能會幫你很大忙，當你輸入少數幾個關鍵詞之後，搜索引擎就會根據你輸入的內容，揣測你的搜索意圖，不僅給你提出建議，還會結合這幾個關鍵詞的語境自動補全搜索關鍵詞，自動匹配相關內容，讓你瞭解到其他人的想法。

如果你不熟悉自動完成功能，可以嘗試一下。^[16]當你在谷歌搜索欄裡輸入「誰是」一詞的時候，谷歌就會自動匹配出其他人經常搜索的內容，比如，自動蹦出「世界首富」等內容。隨意嘗試幾次，你就會發現，人們有時候很想了解另一個性別的人。比如，你輸入「為什麼女性」，谷歌會自動為你匹配出「騙人」「來月經」「穿高跟鞋」等；再比如，當你輸入「為什麼男性」，谷歌則會自動為你匹配出「退縮」「墜入愛河」「撒謊」等。

如果你想在現實世界中窺探一個人對他人的先入之見，這無異於在玩「禁忌遊戲」，但在網絡世界中，卻不存在任何禁忌。為什麼黑人.....喜歡炸雞？為什麼穆斯林.....厭惡美國？為什麼亞洲人.....看起來長得都很像？自動完成功能會自動為你匹配出這類內容。這些例子是

我從谷歌搜索引擎一字不差地照搬過來的。事實上，有一個匹配結果「為什麼白人嘴脣薄？」恰恰是最近一篇研究論文的主題。^[17]這篇論文探討了自動完成功能的雙重作用：一方面，這一功能揭示出了已經形成的趨勢；另一方面，由於谷歌無處不在，也對趨勢起到了塑造作用。這篇論文指出，自動完成功能理應僅僅反映出人們的先入之見，但最終會導致這些先入之見長久存在。這一點並不難理解：用戶輸入關鍵詞之後，卻蹦出了其他人的偏見，很容易影響到用戶的看法。比如，剛剛我輸入了「為什麼同性戀」這幾個關鍵字，隨後，「情侶看起來很像」這幾個字就蹦了出來。我原本不這麼認為，但這樣一來，我反而對同性戀者形成了這種看法。由此看來，搜索引擎不僅會侵犯你的隱私，還會餵你精神鴉片。

如果你輸入一些針對自身的問題，就會從另一個角度瞭解到人性，他人的一切想法都一覽無餘，就像通過浴室鏡去觀察一個人一樣。你可以在搜索欄中輸入「為什麼我的」，那麼谷歌就會根據你的提示為你匹配出一連串讓你頭痛的事情，一些最精彩的就是：「為什麼我的大便是綠色的？」「為什麼我的舌苔是蒼白的？」「為什麼我的尿渾濁？」「為什麼我的陰道癢？」等等。^[18]

必須指出，所有這些問題，可能都是個別搜索者在電腦前坐得太久導致的。但他們在谷歌上搜索之後，自動完成功能就會將他們的想法呈現在他人面前，從而影響到他人的觀念。

因此，我們自己的隱秘想法在無意之間就影響了世界。通過有創意的打字方法、一些變通性的程序和某些算法，我們就能讓一個人的內心獨白被更多人讀到。我們暴露出了自己內心深處某些有害的、荒謬的想法。我們迫切需要大量的搜索數據來窺探人們內心究竟存在哪些有害的衝動。現在，如果在現實世界中講一些種族主義方面的事情，恐怕不會被公眾接受，你想知道的事情與人們所說的事情存在所謂的「社會期望偏差」。而通過大數據，我們知道，人們在虛擬的網絡世界中仍然在談這方面的事情。此外，雖然我們發現潛在的、隱藏的態度的能力是隨著大數據技術的發展而新近獲得的，但很早之前，我們就有能力利用這些態度。由此來看，大數據技術的重要性更加凸顯，能夠讓我們更好地利用和調整自己的態度。我將引用共和黨戰略家李·阿特沃特（Lee Atwater）的話來解釋這一點。^[19]下面是他在1981年以里根政府成員的身份接受政治學家亞歷山大·拉米斯（Alexander P. Lamis）採訪時，在談到共和黨所謂的「南方策略」之際所說的話：

你們從1954年就開始說：「黑鬼，黑鬼，黑鬼。」到1968年，你們再也不能說「黑鬼」了，不然就會適得其反，導致自己遭受損失。所以，你們必須說一些諸如贊成「用校車接送學生」^[20]以及贊成強化州權之類的話。現在，你們要講一講減稅的事情，你們所講的一切都是經濟事務，而這種趨勢導致的一個負面結果就是黑人比白人承受的損失更大。

阿特沃特認為他講的這番話是非正式的，不留記錄，後來卻由於偶然因素而流了出去。後來，當他聽到有人引用他這番話時，他驚訝地反問道：「你確定你引用的話是我說的嗎？」現在，有了搜索數據，就意味著我們不必等待這樣的偶然因素去幫助我們分析人們在種族等話題上的公開言論與真實心聲之間的差距。數據表明，我們的世界正在變得越來越美好，但也表明我們仍有很長的路要走。

我們看一看奧巴馬在2009年發表的就職演講。當時，他講了很多讓人滿懷希望的話，很多人覺得美國已經成為一個「後種族主義」的社會；很多人認為，「後種族主義」並不是一個遙不可及的想法。「後種族主義」這一願景的核心在於將奧巴馬的成功競選拓展到了美國生活的其他角落，說他的勝利證明了在美國人的生活中，種族再也不是一個妨礙成功的因素了。

儘管這一美好的願景有希望成為現實，但谷歌公司的數據科學家、經濟學家賽思·斯蒂芬斯—達維多維茨得出的結論卻是，奧巴馬的種族因素可能導致他在2008年大選中的得票率降低3~5個百分點。損失的不是共和黨的選票，而是民主黨的選票。很多民主黨人寧願把票投給白人約翰·克里，也不願投給奧巴馬。如果奧巴馬的得票率再提高5個百分點，那麼他的得票率將會超過二戰後50%的總統的得票率。如果沒有大數據，我們永遠不可能發現這個結果。這位數據學家還研究了2004—2007年的競選情況，當時，奧巴馬還沒有參與全國性的競選活動。他還根據「谷歌趨勢」分析了美國人在種族問題上的態度。（這樣一來，人們就不至於因為某些人不喜歡奧巴馬本人而對其本人產生誤解。）他使用大數據計算出了各州的「種族敵意指數」，然後將這個指數同奧巴馬最後得到的總票數做對比，並預測出如果一位民主黨的白人處在奧巴馬的位置上，會得到多少選票。（當然，他肯定有足夠豐富的歷史數據來做這項預測。）可以肯定的是，種族敵意指數越高，奧巴馬的得票情況越不樂觀。這位數據學家所說的一番話對這一方法進行了解釋：

考慮一下丹佛和惠靈這兩個地方的情況。^[21]（惠靈夾在俄亥俄州

和西弗吉尼亞州中間。）約翰·克里先生在這兩個地方的得票率均為50%。基於民主黨在2008年的輝煌戰績，奧巴馬先生在這兩個地方的得票率應該達到57%左右。不過，丹佛和惠靈卻表現出了不同的種族態度。按照從高到低的順序排列，在丹佛，種族歧視詞語的搜索率位居全美倒數第四位，奧巴馬在這裡贏得了57%的選票，是符合預期的。但在惠靈，種族歧視詞語的搜索率卻高得多，排在全美正數第七位，結果奧巴馬在那裡只贏得了48%的選票。

從歷史上看，在其家鄉所在的州，總統候選人的得票率會略高於他在美國其他州的平均得票率，大約高出兩個百分點左右。由於種族敵意的影響，2008年，約翰·麥凱恩在其他州的得票率反而高於其家鄉所在州的得票率，這說明他比之前歷屆大選的競選者更容易受到其他州的青睞。他似乎成了美國最受青睞的人。但之所以出現這種情況，只是因為他的對手是一位黑人，而白人選民中間存在廣泛的種族敵意。

在我看來，拳王阿里是最勇敢的美國人之一。^[22]1967年，作為重量級冠軍，正值其職業生涯的黃金時期，他拒絕服兵役，並在媒體上公開發表反對越南戰爭的宣言，震驚了整個美國。美國地方法院以拒絕服兵役的罪名強制收回了阿里的拳王桂冠，吊銷了他在全美各州的拳擊執照，並沒收了他的護照，還禁賽3年半，並判處5年監禁。在強大的壓力面前，阿里並沒有屈服，堅持自己的信仰。他經常出現在各種集會上，甚至電視節目當中，進行反戰宣傳。20世紀60年代末，美國國內的反戰呼聲越來越高，作為反戰人士的代表，阿里也獲得了更多支持。1970年，美國最高法院裁定，恢復阿里的拳手資格，這位經過兩年多修整的前拳王終於重出江湖。難以想象，即便今天的政治領導人也很難具有這樣的魄力，更不用提我們的運動員和名人了。從坎耶·韋斯特（Kanye West）到格林·貝克（Glenn Beck），從雷切爾·瑪多（Rachel Maddow）到莎拉·佩林（Sarah Palin），在面對種族主義時，我們只是看到他們的憤怒，卻沒有人願意像阿里那樣為了堅持信仰而犧牲自己的利益。對於阿里在越南戰爭上的立場，我們每個人都會有自己的看法。我本人是一位老兵的後代，我知道不止一人不認同我的看法。即便通過今天的搜索數據，我們也不難理解阿里當時為什麼會持有那種觀點。正如他所說的那樣，「沒有任何一個越共叫我黑鬼」。或許，他是正確的。但想象一下，如果當時谷歌已存在的話，美國人會在搜索欄中輸入什麼信息。想象一下那些歲月，一個黑人在自己的國家遭遇了多少不公正的待遇。

接下來美國人在種族問題上的態度會何去何從還有待觀察。儘管存

在上述種種不容樂觀的情況，奧巴馬畢竟當上了美國總統，有一些事情令人失望，也有一些事情令人深受鼓舞，其中一件事就是，沒有證據表明奧巴馬總統在2012年的選舉中遭到了種族偏見的傷害。那時，他是一個名人，人們在更大程度上知道他是「巴拉克·奧巴馬」，而不是「一個黑人」。在本書的整合數據中，缺失了個人層面的數據，而這些廣泛的社會力量對個人的影響往往是非常微妙的。以前一章的數據為例。OkCupid網站上的許多黑人用戶也擁有過愉快的經歷，如同其他人一樣，黑人也得到過約會機會，也遭到過拒絕，只是總的來說，後者多於前者。如果你從每一個人的角度去分析問題，那麼個人的經歷就會變得太渺小，差異性太大，使你很難根據個人經歷去得出某些關於「種族主義」的結論。個人的不幸遭遇可能是由於膚色引起的，也有可能是由其他方面的因素引起的。如果沒有大數據，你看到某個人因為奧巴馬當選總統而氣得面紅耳赤地去上網搜索「黑鬼笑話」，那麼這種做法就非常有趣、非常可笑。但現在有了大數據，你看到有成千上萬人去搜索，而他只是其中之一，可能就不那麼有趣了。當你看到這些隱祕的態度仍然產生很大的影響，甚至在公共生活中產生很大影響時，可能也沒那麼有趣了。沒有大數據，我們只能看到個別的情況，而有了大數據，我們就能看到所有人的情況。正是由於這個原因，這類數據是非常必要的。大數據讓我們看到了無數人的隱祕選擇，讓我們看到了自己必須面對的嚴峻事實。

我知道有些人只讀好書。所謂「好書」，我的意思是朋友、老師、書評人、亞馬遜等推薦的書。這不難理解，閱讀過程是緩慢的，時間又無比寶貴，為什麼要冒著風險自己去選書呢？但這並不是我的風格。我喜歡歷史，去書店時，直接到歷史類書籍的架子上隨機抓幾本。作為一個讀者，我的確讀到過不少沒有意義的東西，但也有很多有意義的發現。我讀過很多關於拿破崙的書。在偶然發現的眾多好書中，《美國人民的歷史》（*A People's History of the United States*）是我最喜歡的書之一。是的，我現在知道這是一部經典之作，但這並不能改變這一事實，即在我從書架上把這本書拿下來之前從未聽說過它。谷歌圖書對其進行了很好的描述，將其稱為「一部自下而上的美國通史」。大多數書籍講的都是領導人 and 大事件，而這本書卻為我們講述了歷史上的家庭、商店、農場、工廠和平民內心的憂慮。但事實上，雖然我愛這本書，雖然它顛覆了美國課堂上的歷史教科書，本書作者霍華德·津恩（Howard Zinn）只能跟我們分享他自己的見聞與感想，只能告訴我們他人大聲講出來的事情。他無法看出人們埋藏在內心深處的想法。面對古巴導彈危

機，面對百無聊賴的戰壕，面對避孕藥帶來的性解放，人們內心究竟在想什麼？歷史沒有記錄人們內心深處的歡樂與痛苦，但如果我們當時擁有現在這些大數據的話，會出現什麼情況呢？我們對自身、對歷史的理解與認知將會變得更加充實。

[1] 關於蓋洛普起源的資料，摘自維基百科中「Gallup (company)」這一條目的內容。

[2] 我在本章正文和腳註中都提到過，正是谷歌公司數據科學家斯蒂芬斯—達維多維茨 (Stephens-Davidowitz) 想出了利用「谷歌趨勢」去跟蹤各種禁忌話題的主意。本章的靈感來源於他於2012年6月9日在《紐約時報》上發表的「How Racist Are We? Ask Google」一文以及他2013年的哈佛大學博士論文「Essays Using Google Data」，獲取鏈接為：<http://nrs.harvard.edu/urn-3:HUL.InstRepos:10984881>。本章稍後會提到奧巴馬在大選期間因為種族問題而失去了多少選票，我在這一點上參考了達維多維茨的研究成果。關於「nigger」這個單詞在不同時期的使用情況，以及本章提到的其他利用「谷歌趨勢」取得的發現，都是我根據戴維德威茨的建議，通過自己的研究得出的結論。雖然達維多維茨現在供職於谷歌公司，但我必須強調他在研究搜索數據時，依靠的都是公開的、匿名的數據來源，而不是依靠工作便利獲取的私人搜索內容。我自己在研究搜索內容時，依靠的也是公開的、匿名的數據來源，即Google Trends: google.com/trends。

[3] 這裡的細節參考了維基百科上的「social desirability bias」條目的內容。

[4] 我首次關注這個效應是在2008年美國大選期間。當時，很多權威人士都想知道這種效應會對奧巴馬在選舉日當天的民調有什麼影響。這一段提到布萊德利敗選的細節，參考了維基百科上「Bradley effect」條目的內容。

[5] See Nick Bilton, 「Google Search Terms Can Predict Stock Market, Study Finds,」 New York Times Bits blog, April 26, 2013. See also Casey Johnston, 「Google Trends Reveals Clues About the Mentality of Richer Nations,」 Arstechnica, April 5, 2012, arstechnica.com/gadgets/2012/04/google-trends-reveals-clues-about-the-mentality-of-richer-nations/; and Tobias Preis et al., 「Quantifying the Advantage of Looking Forward,」 Scientific Reports 2, no.350 (2012), doi:10.1038/srep00350.

[6] Google Flu was first developed in the paper 「Detecting Influenza Epidemics Using Search Engine Query Data,」 by Jeremy Ginsberg et al. in Nature 457 (2009):1012—14, doi:10.1038/nature07634. Recently, Flu's efficacy has been found wanting: see Kaiser Fung, 「Google Flu Trends' Failure Shows Good Data > Big Data,」 Harvard Business Review Blog Network, March 25, 2014.

[7] 這一點引用了達維多維茨在「How Racist Are We? Ask Google」一文中提出的數據。

[8] 「谷歌趨勢」根據某個單詞或短語的搜索量編制了一個具有相應比例的指數，來表示這個關鍵詞的搜索頻率。上述幾個城市中，nigger一次的指數相差在10%以內。我沒有將nigga一詞統計進來，因為這個詞通常是與rap，即說唱藝術放在一起搜索的。到目前為止，與nigger一詞聯繫在一起的最常見的搜索詞是joke，意為「笑話」。我在統計與種族主義相關的搜索詞時，採用的是谷歌公司數據科學家賽思·斯蒂芬斯—達維多維茨 (Seth StephensDavidowitz) 創造的統計方法。作為網站管理者，他能接觸到用戶輸入的數據。他根據獲取的這些「內部消息」寫道：「nigger一詞的搜索中，很大一部分都是搜索與非洲裔美國人有關的笑話。」他在研究過程中使用的都是公開的、匿名的數據。

[9] 根據谷歌趨勢發佈的美國搜索指數顯示，從2004年1月到2013年9月，apple pie的指數為25，nigger的指數為32，比前者多出近30%。

[10] 在我的Twitter語料庫裡面，nigga與nigger的比例比二者在谷歌趨勢上的比例高出30倍。換言之，在Twitter上，nigger的出現頻率只有1/30。

[11] 最低點之所以出現在這個時期，並不是因為所有人都去度假了。pasta（意大利麵）、pizza（比薩）、family（家庭）和truck（卡車）之類的中性詞的搜索次數在2008年全年都保持穩定。

[12] KKK，意為「3K黨」，是美國最悠久、最龐大的種族主義組織，奉行白人至上主義，是美國種族主義的代表性組織。——譯者注

[13] 這個數據摘自達維多維茨發給我的郵件。

[14] 這個數據摘自達維多維茨的「How Racist Are We? Ask Google」一文。通過谷歌趨勢，也能直接得到證實。

[15] 達維多維茨在發給我的電子郵件上說，這些種族歧視的詞在Twitter、OkCupid和谷歌搜索上使用的普遍性要低得多。

[16] 谷歌自動完成功能依賴的算法可以說是非常神祕的，關於其運作機制，幾乎查不到任何明確的資料。searchengineland.com的珍妮·沙利文（Danny Sullivan）曾經提出過大框架上的分析，但也不算透徹，具體可以參考searchengineland.com/how-google-instant-autocomplete-suggestions-work-62592。由於自動完成功能與每個人的搜索記錄有關，因此，如果嘗試的話，大家得出的結果可能是不一樣的。在嘗試時，請記得像我一樣使用Chrome瀏覽器的無痕瀏覽功能，這樣一來，谷歌就不會把你之前的搜索記錄納入考量範圍了。如果你用的是Safari瀏覽器，那麼請開啟私密瀏覽模式。

[17] See Paul Baker and Amanda Potts, 「‘Why Do White People Have Thin Lips?’ Google and the Perpetuation of Stereotypes Via Auto-Complete Search Forms,」 Critical Discourse Studies 10, no.2 (2013): 187—204.

[18] 下面這一長串問題是肖恩·馬太（Sean Mathey）給我提出的建議，當時我們剛剛結束一次露營，開著車回家。那次露營的時候，我們玩了一種名叫「戒指連牌」的魔術。

[19] See Rick Perlstein, 「Ex-clusive: Lee Atwater's Infamous 1981 Interview on the Southern Strategy,」 The Nation, November 13, 2012, thenation.com/article/170841/exclusive-lee-atwaters-infamous-1981-interview-southern-strategy.Original quote from Alexander P.Lamis's book The Two Party South (New York: Oxford University Press, 1984), via Wikiquote's 「Lee Atwater」 entry.

[20] 指贊成為平衡黑白學生比例用校車接送外區兒童上學。——譯者注

[21] 摘自達維多維茨的「How Racist Are We? Ask Google」一文。

[22] 早在1999年，我就讀過戴維·雷梅尼克（David Remnick）為拳王阿里撰寫的傳記《世界之王》（King of the World）（New York: Random House, 1998）。自那之後，我就一直非常欽佩阿里。為了確認他抗議越南戰爭的情況，我參考了維基百科的內容。關於他提及的那句名言（「沒有任何一個越共叫我黑鬼。」），我在本書引用的其實並不是完完整整的原話，但我引用的這個版本比較凝練，更加為人熟知。他的原話是：「我的良心不允許我為了強大的美國舉槍射向自己的兄弟，或是一些深膚色的人，貧窮、飢餓、深陷泥潭的人，對他們開槍是為了什麼？他們從沒叫過我黑鬼。他們從沒對我用過私刑。他們沒對我放過狗，沒有剝奪我的國籍，沒有強暴殺害我的父母……我為什麼要殺他們？我怎麼能射殺那些窮困的人？你們還是把我關押起來吧。」因為我在本書引用的那句話在實質內容上沒什麼區別，更加簡短，知名度更高，所以我還是決定採取那個簡短的版本，而沒有完整引用原來的版本。如果想聽一聽阿里親自講出這些話（即比較長的版本），請上YouTube搜索「Muhammad Ali on the Vietnam War-Draft」，鏈接為：<https://www.youtube.com/watch?v=HeFMMyrWlZ68>。在視頻裡，他似乎是在一次比賽結束之後接受採訪時說這番話的，語調緩慢，顯然是經過認真思考的。兩年之後，他又就同一個主題發表過一次比較流利的講話，請搜索「Muhammad Ali Interview with Ian

Wooldridge（1969）」，鏈接為：https://www.youtube.com/watch?v=dLam_GiQ2Ww。

第九章 憤怒的時代

新年前夜，17歲的北卡羅來納州女孩薩菲亞·納瓦茲（Safiyyah Nawaz）無聊地坐在沙發上，等待著水晶球的降落。她在Twitter上發了一個愚蠢的笑話：

\$afiyyah @safiyyahn

現在，這個美麗的地球2014歲了，好神奇。^[1]

她獲得了1.6萬次轉發，幾乎都是在她發帖後24小時之內轉發的。為了對比起見，我們看看其他人轉發的情況。凱蒂·佩裡（Katy Perry）在Twitter上有4900萬名粉絲，她發的「新年快樂」的祝福只勉強得到了1.9萬多次轉發。歌手Lady Gaga發出了一段粉絲期待已久的視頻，也只是獲得了2萬次轉發。^[2]薩菲亞·納瓦茲並不是一顆冉冉升起的世界巨星，她的情況也算不上自命不凡的暴發戶利用Twitter挑戰文化秩序。如果你沒聽說過薩菲亞，這很正常，因為她只是北卡羅來納州的一個高中生，而她發的笑話，也就是上面的帖子，卻引爆了Twitter。

一開始，網友們只是發一個撓頭的表情，表示疑惑不解，不確定她說的話是不是認真的。但如果你多看一些回覆，就會發現網友越來越過分，逐漸開始攻擊、謾罵薩菲亞。這些荒謬的回覆是一種普遍現象，你看一下轉發數量就知道了。事實上，你會發現，在網絡世界中，人很容易變成暴民。在很短的時間內，網友的回覆從揶揄式的大笑（LOL，laugh out loud）變成了「噢，我的上帝」（OMG，Oh, my god），繼而變成了「他媽的」（WTF, what the fuck），到最後就出現了下面這類東西：

Cocaine Burger @Cocaine_Burger

@safiyyah Kill yourself（死去吧你）

Rick Huijbers @HARDEBAKSTEEN

@safiyyah Kill yourself you stupid motherfuck（死去吧你，蠢豬）

正如著名的商業博客Gawker在其報道中所說的那樣，在短短幾分鐘的時間內，網友回覆從「傻瓜」變成了「蠢豬」。考慮到網友的回覆充

斥著暴力，作為一個17歲的孩子薩菲亞對這種情況的處理方式已經算是非常好了，後來，她對網友憤怒的回覆進行了如下完美的概括：

\$afiyyahn @safiyyah🐦

young folks these days b really passionate about the tru age of the earth（這些天，年輕人真的非常熱衷於探討地球的真實年齡。）

她可能沒有意識到，還有一個名人和她遭遇了同樣的事情。就在她發出那條笑話15分鐘之前，著名的喜劇演員娜塔莎·賴格羅（Natasha Leggero）正在時代廣場上和主持人卡森·達利（Carson Daly）錄製節目，兩人打趣地談到了美國知名罐裝意大利麵品牌SpaghettiOs在紀念珍珠港日時的一次公關活動。^[3]這個品牌之所以飽受非議和批評，是因為它鼓勵人們通過購買罐裝意大利麵來紀念陣亡的將士。娜塔莎說：「珍珠港事件為數不多的倖存者只能嚼得動這類食物了，但這個品牌卻拿這種食物來取笑他們，簡直糟糕透了。」

然後，主持人與嘉賓談笑風生地轉移到了其他話題，但娜塔莎卻無意之間觸碰到了網民們高度敏感的神經，點燃了他們心中的怒火。自以為秉持正義的網友們很快對她發動了一輪又一輪的攻擊謾罵。娜塔莎後來在Tumblr上貼出了她收到的一些回覆，很多是下面這樣的：

Mike Oswald @SDPStudio🐦

@natashaleggero What a vile whore you are.（你個無恥的婊子。）

Mark Tichenor @hotrod607🐦

@natashaleggero Fuck you, you disrespectful cunt.（滾，你個討厭的女人。）

如果哪一天互聯網死掉了，我們要給它寫個墓誌銘的話，那麼我覺得用下面這個回覆最好不過了：

Chris McAllister @macdawg22🐦

@natashaleggero you're a stupid ignorant whore.（你是一個愚蠢無知的婊子。）

我之所以對這兩個人物的遭遇給予特別關注，是因為我的一個同事賈斯汀·薩科前不久也遭遇了類似的事情。她是OkCupid網站母公司InterActiveCorp（IAC）公共關係部的主管。12月20日，她在倫敦希思羅機場準備轉機到約翰內斯堡。她登上飛機，坐到座位上，拿出手機，

在Twitter上發了下面一個帖子：

Justine Sacco @justinesacco

Going to Africa.Hope I don't get AIDS.Just kidding.I'm white!（要去非洲了。希望不要染上艾滋病。開玩笑啦。我是白人！）

然後，她關掉了手機。與前面講過的兩個案例相比，她的帖子不太像個笑話，充其量只算是道出了白人的特權，挖苦了一下黑人。但這是一種拙劣的做法，很容易讓人想到種族主義。對於她這種傻乎乎的做法，網友們一開始只是搖搖頭，但很快演變成了激烈的人身攻擊。她遭到了威脅和侮辱，但網友的攻擊對象不僅僅是她自己，她家人的照片也被人肉出來，在網絡上瘋傳，甚至連家人的行蹤也被傳開了。^[4]有些男性網友甚至給她的侄女們打電話，威脅說要強姦她們。約翰內斯堡機場也聚集了一大批人，等著她的飛機降落。在飛行期間，她無法做出迴應，從而助長了網友的氣焰。大概飛行了一半距離時，「賈斯汀到了嗎」（#HasJustineLandedYet）這個標籤就被創造出來，迅速成為Twitter上最熱門的話題。在谷歌上搜索她的名字，搜索結果開始自動給出她的航班號和降落時間，因為這就是當時人們搜索的內容。搜索算法再次舉起了一面鏡子，照出了人們內心的共同想法。賈斯汀在空中飛行的11個小時裡，科技博客Valleywag首先挑選出這個帖子進行公示，接著主流媒體《紐約時報》和CNN（美國有線電視新聞網）加入報道，她的僱主IAC集團迴應表示震驚，無數網民留言批評，一個個怒氣衝衝，迫不及待地等著她降落。^[5]她下飛機後不久便遭到了解僱，她簡單的幾句話就輕而易舉地毀掉了自己的生活。

RonGeraci @RonGeraci

It's like 2 million people are waiting for her with the lights off to see her expression as the earth explodes.（似乎200萬人都在等她，關掉燈，看看她在地球爆炸時的表情。）

I'm Gary @noyokono

#HasJustineLandedYet People haven't eagerly anticipated a plane landing this much since Amelia Earhart.（#賈斯汀到了嗎？自阿梅莉亞·埃爾哈特之後，人們就從沒有如此熱切地期待飛機降落。^[6]）

V.Hussein Savage @Kennymack1971

Aw hell....lemme finish this work grab a 6 pack and some BBQ wings.

It's about to be on...

#HasJustineLandedYet（哦，幹完活兒，練練腹肌，吃點烤雞翅，好戲就該上演了.....#賈斯汀到了嗎？）

他們的「獵物」本來只有幾百個粉絲，也沒有什麼公開資料。我不是非常瞭解賈斯汀，但我和她有過一段愉快的共事經歷，但看到網民們給賈斯汀造成巨大的痛苦和恐懼之後的興奮和激動，真的令我很厭惡作嘔。

就像傻瓜一樣，我去Facebook上發洩了一下自己的情緒。我的帖子剛發上去不到10分鐘，一位15年沒有交談過的熟人發了一個評論貼，說「她的父親是個億萬富翁」，言外之意就是她的毀滅也是罪有應得的。^[7]後來，我在Facebook上直接把他從好友列表裡刪掉了。賈斯汀的父親肯定不是一位億萬富翁，這是無中生有，純粹是謠言。^[8]這就像在石刑現場遇到了一群暴徒，當你試圖把人們拉走的時候，突然發現一個熟人，於是暗自慶幸終於找到一個可以理論的人了，結果對方卻睜大雙眼，對你喊道：「老兄，看看所有這些石頭。」

當我閱讀關於這類事件的評論時，一次又一次地想起石刑的場景。古代宗教選擇石刑作為執行死刑的手段絕不是巧合。石刑的獨特之處就在於劊子手不是單個的人，而是所有人，施行懲罰的是一個集體，所有人都在那裡朝著受害者扔石頭，誰也說不清楚究竟誰扔的那塊兒石頭造成了致命一擊。^[9]在一個充滿敵意的世界裡，新興部落需要通過戰鬥來維繫自身的生存，維護自己信奉的神靈，維護部落成員的凝聚力。如果採用其他手段來施加死刑，就會讓劊子手產生嚴重的內疚感。相反，石刑造成的是集體內疚，每個成員心裡都存在一定的內疚感，反而減輕了內疚感，在施加懲罰的同時也不至於削弱部落的凝聚力。除了石刑之外，還有什麼更好的辦法嗎？

在賈斯汀的案例中，來自三個大洲的人們聚集在一起，共同執行「石刑」，毀掉了賈斯汀。看一看這些網友在Twitter上發佈的個人簡介，就會發現他們給自己貼上了各種各樣的標籤，包括遊說者、憤青、亞斯伯格症患者、領導者、自然愛好者、博主、短吻鱷、父親、作家、搖滾音樂迷、不完美的基督徒、放射科醫師、流行文化愛好者、海的女兒、風的姐妹等。這些人之間共同之處僅僅侷限於他們具有一個共同的抨擊目標，都在Twitter上給自己貼了個標籤，並且都達到了自己的目標：賈斯汀丟了工作。美國的新聞聚合網站BuzzFeed把賈斯汀的頭像放到了首頁，並配上一個大笑表情以示嘲諷。

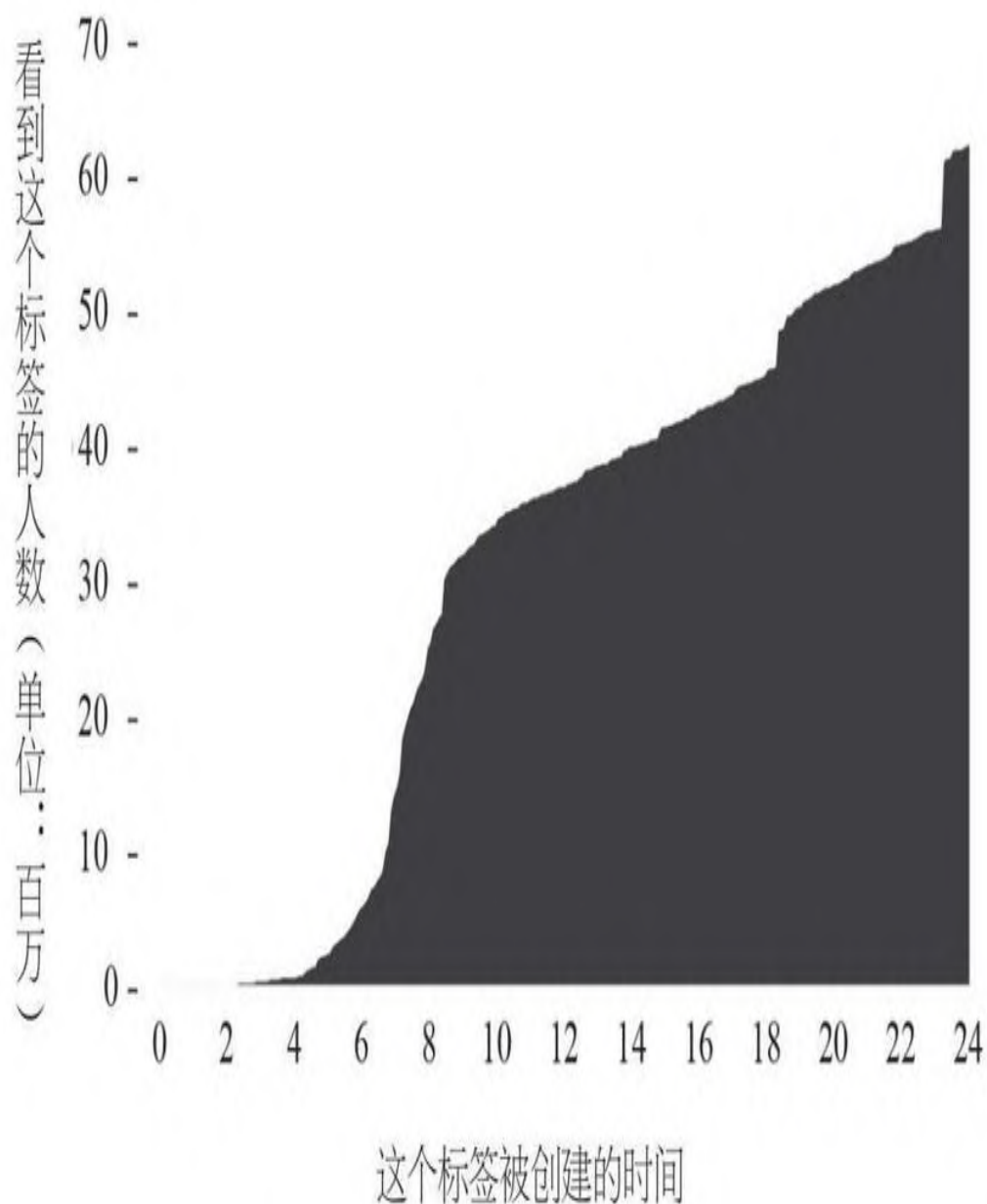


圖9—1 「賈斯汀到了嗎」這個標籤引發的關注

社交媒體的覆蓋範圍非常廣，從而放大了網友的力量。^[10]薩菲亞那個帖子發出後不到24小時，指責者就飆升到了740萬人，6200萬人看到了「賈斯汀到了嗎」（#HasJustineLandedYet）這個標籤（見圖9—1）。

在圖9—1中的這條曲線下面，雖然很多人看到了「賈斯汀到了嗎」這個標籤，但並非每一個人都讀到了最原始的帖子，也並非每一個人都

很在意這件事情，但的確有很多人會讀到原始帖子，也非常在意。從某種意義上來講，每個看到這個標籤的人都是這一事件的見證者。很多人看到這個標籤後都會在好奇心的驅使下追溯原始的帖子，看看究竟是怎麼一回事：

Sir QwapQwap @BeardedHistoria 🐦

我Twitter主頁上前20個帖子裡幾乎都有「賈斯汀到了嗎」的標籤，我這個Twitter迷肯定錯過了一些信息。

值得指出的是，這種不可思議的關注量應該成為社交媒體的一個恥辱。這在印證網絡力量十分強大的同時，也暴露出了網絡力量多麼虛偽，因為賈斯汀的案例為我們揭示出了人類社會中的重大問題，比如艾滋病、種族歧視以及後殖民時代非洲長期存在的貧困問題。我們本應該為人類社會存在這些問題而感到恥辱，然而，Twitter上幾乎沒有網友關注這些問題，他們不約而同地將注意力用在了批評賈斯汀上。Twitter沒有為解決這些問題做出任何貢獻。

我們可能認為用活人獻祭之類的事情只存在於野蠻殘酷的遠古時代，只有在關於寺廟和世界末日的電影裡才能看到類似的場景。其實不然，現代社會的很多動物仍然存在這種本能，只不過在漫長的進化過程中深深地根植於其內心世界，沒有遠古時代表現得那麼明顯。比如，當食物稀缺時，獅子會殺死自己的幼崽將其吃掉；魚類會吃掉自己產的卵。在人類身上也存在類似傾向，比如，在胞胎妊娠的情況下，子宮有時會自發地選擇將一個胎兒血液輸送給另一個胎兒。為了眾人利益而犧牲個別人利益的做法由來已久，可能與生命的歷史一樣悠久。我們在前面提到了網絡攻擊也具有這樣的屬性，不過值得慶幸的是，人們的手上並沒有沾上真正的鮮血。在Twitter上看帖子的時候，你知道這只是人們在網絡空間表現出來的一種偏好，在現實世界中未必會表現出來。得益於互聯網技術的發展，我們現在終於可以藉助大數據對其進行直觀的研究，這是有史以來的第一次。過去，由於缺乏互聯網的協助，社會科學家們為了研究消極思想的形成原因和傳播方式，不得不依靠傳統的問卷調查方式，耗費了相當多的時間和精力。現在，互聯網為他們提供了無限的原始資料以及強大的追蹤機制。海洋生物學家可以將一些高科技的標籤貼到鯊魚身上，^[11]利用聲呐及全球定位系統技術瞭解它們的行蹤，減少它們對人類的威脅。^[12]我在前面所舉的三個例子都是切實存在的，不是謠言，很多情況下，民眾就是通過這些案例中的方式來宣洩自己的憤怒。認真研究一下關於謠言的科學，可以幫助我們理解娜塔

莎、薩菲亞和賈斯汀到底遭遇了什麼以及為什麼會出現這些遭遇。

我們在本書伊始就提到了謠言。^[13]挪威、埃及和希臘在遠古時期的眾多神靈中，肯定有一個與謠言相關。《舊約·箴言》裡面就包括很多關於謠言這個主題的至理名言，其中有一句話是：「藐視鄰舍的，毫無智慧，明哲人卻靜默不言。」《聖經》裡還有一句非常著名的至理名言就是：「若不想被人議論，自己最好不要議論別人。」有幾份資料表明，古羅馬人信奉的神靈中，有一個被稱為「謠言女神」。她長著翅膀，有100隻眼、100張嘴，到處傳播謠言，中傷他人，破壞人與人之間的感情。不過，對於這一點，我不太確定，所以不進行更加深入的闡述了。

進化生物學家相信，之所以會出現謠言，根源在於我們的祖先需要通過語言交流去了解其周圍環境。他們的理論是，如果古人必須弄清楚某件事情的真實性，那麼語言為他們提供了一種調查和討論的方式。無論人們說的話是真是假，都會逐漸傳播開來，經過多人的轉述，便形成了謠言。所謂謠言，本質上是一個群體對某件事情真實性的猜測。謠言逐漸變成了人們建立社會關係、積累社交資本的一個途徑。誰分享和傳播了謠言，誰就能從中得到一定的社會地位，尤其是關於重要人物的謠言，因為關於權勢人物的信息本身就是一種權力的形式。

但社交媒體的出現給這種局面帶來了多重改變。第一改變是，社交媒體給我們提供了一些可以用來判斷一個人社會地位的指標，比如粉絲數量、轉發數量、點贊數量等。如果你是某條新聞的第一個傳播者，說了一些特別鞭辟入裡的話，獲得較多的轉發，那麼你的粉絲們就會為你的智慧鼓掌點贊。現在，你通過分享信息建立起來的社交資本是顯而易見的。事實上，你的社交資本就體現在你電腦屏幕上緩緩增加的那個小小的數字上，這個數字就在你的眼前。傑西·辛格爾（Jesse Singal）在《波士頓環球報》上發表的一篇文章中談到了傳統上人與人之間傳播流言蜚語的動機。他認為，人們在傳播謠言時懷有特定的動機，他們更注重向誰傳播謠言，而不是注重謠言中的主人公是誰。這句話也適用於那些在Twitter上傳播謠言的人們。在傳播謠言的問題上，互聯網製造了空前廣泛的受眾。

第二個變化是，互聯網使每個人都變成了公共人物。之前，地位高的人往往都是酋長，後來是名人和總統；但現在，互聯網技術展現了其非凡的影響力。藉助互聯網，任何人都可能在一夜之間聲名鵲起，當然，也可能一夜之間聲譽掃地。有些人像傳教士般地說互聯網會讓人們

獲得更多的能量，這是最不喜歡的言論。講這些話的人和他們的投資者的確會因為互聯網的發展而獲得更大的能量，但對於更多的普通人來講未必如此。不過，這些陳詞濫調般的言論也具有一定的道理。人們遭到中傷的風險大為增加。互聯網為傳謠者提供了造謠和中傷他人的工具，一個人可能會遭到100萬個人的中傷。

與傳統的信息傳播方式相比，網絡傳播有很多獨特之處，比如不同步、匿名、容易讓人逃避現實、缺乏核心權威等。這些特性使人們無法駕馭信息的流動，淡化了傳遞信息時的責任感，有利於促進人與人之間的交流。但從另一個角度來講，這也讓互聯網交流變得令人望而生畏，因為人們在虛擬的網絡世界裡想怎麼做就怎麼做，想說什麼就說什麼，卻幾乎不用承擔什麼後果。作為最早研究這種現象的學者，新澤西州萊德大學心理學教授約翰·蘇勒（John Suler）將這種現象定義為「網絡抑制解除效應」。^[14]這一理論是指，在虛擬的網絡空間中，沒有人知道你是誰，你擺脫了自身的身份，而身份對自己行為的限制也會消失。在線漫畫網站Penny Arcade對這種現象進行了更加形象的概括，它提出了「網痞理論」，即認為「網痞=網民+隱蔽性+受眾」。

但與傳統信息傳播媒介相比，網絡的獨特之處並不在於網絡言論的惡毒，也不在於網絡人物是匿名的。互聯網在這兩個方面並沒有帶來什麼革命性的變化。比如，卡車司機收聽的無線電廣播就充斥著種族主義的謾罵。^[15] ^[16]在來電顯示技術出現之前，《搗蛋雙寶》的主人公在長達數十年的時間裡利用電話的匿名性多次欺騙和謾罵他人。^[17]即便到了今天，仍然有人利用無線電廣播的匿名性相互攻擊。像這樣的消極信息具有漫長的歷史。^[18]與無線電廣播和電話相比，互聯網真正的獨特之處就在於它給我們帶來了強大的追蹤機制，我們終於可以採取一些行之有效的措施加以迴應了。比如，我們曾經在第七章討論過，微博客Tumblr制定了一系列規則來限制與大腿間距有關的色情圖片。從某種意義上來講，這類限制措施就是人們利用互聯網技術來回應消極信息的一個特例。現在，我們可以確定某個消息最早是誰說出來的，可以確定某個人說了哪些話，確定某個人說某句話的時間，甚至連人們交流時所處的經緯度都能確定下來。正如我在前面指出的那樣，到2015年，人們在Twitter上打的字比人類有史以來印刷的書籍包含的字數還要多。因此，我們面臨的一個非常重要的問題就是如何管理這些無休止的言論。

信息追蹤機制的最大受益者是政府。目前，人們已經可以利用數學模型預測武裝衝突的結果，比如，衝突的持續時間、最終贏家以及死亡

人數等。近來的模型已經逐漸適應了游擊戰，因為今天的戰爭很多都是游擊戰。但在武裝叛亂爆發之前，往往先發生非武裝性的社會動盪。與這類動盪有關的信息往往是通過社交媒體傳播的，組織者甚至通過社交媒體開展協調工作。^[19]現在的運動已經被數字化，吸引了研究人員的關注。

麻省理工學院的彼得·格洛（Peter Gloor）將西方的運動作為研究對象，開發出了一套可以追蹤抗議者情緒變化的軟件。他將這款軟件命名為《禿鷲》，似乎很多依靠政府撥款的項目都是以這個名字來命名的。這款軟件首先確定抗議群體的核心特徵。為此，它會先篩選博客、Facebook以及其他社交媒體上的資料，按照抗議者的粉絲數量對其進行分類，粉絲數量越多，則其影響力越大，再通過數學算法可以確定影響力最大的幾名抗議者。這個過程非常類似於我們之前用很多線條和節點描述夫妻雙方的社交網絡，然後通過數學算法確定下來一些最重要的節點。確定了最重要的幾名抗議者之後，這款軟件就會自動追蹤他們的後續言論。格洛發現，如果一個運動的核心人物在選擇詞語時總是選擇積極的措辭，那麼這個運動正在迅速發展。然而，當積極的、激烈的、煽動人心的言論向消極措辭轉變時，也就是「不」「從不」「蹺腳」「我討厭……」「傻瓜」之類的詞開始頻繁出現的時候，這個運動就開始走向頹勢了。正如《經濟學人》雜誌所描述的那樣，如果示威者埋怨自己所參與運動中的某些人為笨蛋或者出現偷啤酒等不當行為，就預示著整個事態正在接近尾聲。

這項技術不僅可以用來執行暗中跟蹤監視的任務，還能提前破解抗議和動亂者的真正目標。比如，《禿鷲》這款軟件就曾經通過類似的文本分析來判斷哪些埃及城市最容易受到以色列的邊境突發事件的影響，並且能判斷一個乾旱的國家哪些地區最容易出現飲用水危機。

如果一個軟件要跟蹤某條信息的傳播路徑，那麼它不僅需要跟蹤信息本身，還要看到接觸這個觀點的人，看看哪些人最容易受其影響。這些軟件必須能夠發現哪些想法得到了廣泛接受，哪些得到了大量轉發以及信息背後的推手。我們在前面講過，消極思想不是互聯網的專利，同樣，轉發或轉述他人的想法也不是互聯網的專利。電視和無線電廣播也能讓人們轉述他人的觀點。且不提Twitter，即便在美國在線（America Online）出現之前，很多電視節目和無線電廣播節目講到的一些話就變得非常流行，被人們一遍又一遍地轉述。比如，著名廣播談話節目主持人拉什·林堡（Rush Limbaugh）的鐵桿粉絲們就自稱為「應聲蟲」

（Dittohead），他們總是一遍又一遍地重複、轉述林堡說過的話。然而，不可否認的是，與電視和無線電廣播相比，互聯網技術讓轉述、轉發的過程變得更簡單了，也更加容易追蹤信息的傳播路徑，人們只需要輕輕地點擊一下Like（贊）、Recommend（推薦）、Reblog（轉發微博）或Retweet（轉發Twitter）等按鈕就行了。請記住：Twitter上每天的帖子多達5億條，27.5%的帖子都是轉發的，人們只是傳播了別人的思想。[\[20\]](#)

Facebook的數據分析小組調查和分析了2009年醫保改革大辯論期間一條狀態更新的演變情況。這條狀態更新的内容是：

任何人都不應因無力承擔醫療費用而死去，任何人都不應因患病而破產。如果你同意，今天就用這句話更新自己的狀態。[\[21\]](#)

據統計，共有47萬人一字不變地用這句話更新了自己的狀態，這句話還催生了121605條變體，而這些變體本身又得到了80多萬次轉發。如果有人覺得這句話沒有很好地道出自己的真實心聲，就會對其做出小小的改動，於是，不同版本的變體就逐漸傳播到不同的社交圈裡。如果你將這些用戶的帖子與其政治傾向做個對比（—2.0表示最自由的政治傾向，+2.0表示最保守的政治傾向），你就會發現很多有趣的現象（見表9—1）。你會瞭解到美國各個政治派系的情況，包括左派、右派以及不參與爭辯的中間派。此外，你還會看到一個人的政治信仰對其言論的影響。在頂端與底端的兩類人採用的句子結構是一樣的，但其表達的政治傾向卻截然相反。

表9—1 用戶帖子及其政治傾向

任何人都不应……	发帖人的政治倾向
……因无力承担医疗费用而死去……	-0.87 较自由
……因还不起房贷而被冻死街头……	-0.37
……因买不起猎枪而死于僵尸之手……	-0.30
……因患癌症担心明天将会死去……	-0.02
……因负担不起酒钱而没啤酒喝……	+0.22
……因为政府插手医保而死去……	+0.88
……因为奥巴马对医疗资源实行定量配给而死去……	+0.96
……因为政府征税和开支而破产……	+0.97 较保守

1950年，在電視機時代即將來臨之際，美國政治科學協會曾經呼籲美國政治出現一些分化，因為當時兩黨的各项主張過於相似，導致選民難以清楚地予以區分，難以做出明確的選擇。[\[22\]](#)後來，這個協會的目標實現了，但隨著分化日益嚴重，很多人也開始感到後悔。60年後的今天，分化程度已經空前嚴重，你仔細研究一下人們的言論就能發現這一點。無論是國會還是書籍（可以通過谷歌圖書來分析），都一遍遍地重複著黨派色彩非常鮮明的言論。這種現象與當前的政治僵局具有高度相關性。事實上，當前的政治僵局達到了有史以來最嚴重的程度，美國選民幾乎在每一個議題上都會發生分歧。如果說有什麼事實是我們都一致認同的，那麼這個事實就是「我們存在分歧」。

賈斯汀的帖子發出之後，我在Facebook上分析了輿論動向，結果更

加清楚地認識到了美國人的分歧是多麼嚴重。我看到了一篇文章的鏈接，這篇文章出自breitbart.com，這個網站是以茶黨一位極具煽動性的人物——安德魯·布賴特巴特（Andrew Breitbart）的名字命名的。這篇文章裡的很多內容有失偏頗，令人遺憾，但作者卻指出網友們對賈斯汀的帖子著實反應過度了。在當時網友眾口一詞地討伐賈斯汀的情況下，能夠像布賴特巴特這樣冷靜分析的人並不多。在賈斯汀事件之前，我一度認為不分青紅皂白地表達憤怒是右派人物的一個惡習，因為我聽到了右派人士發表的很多荒謬言論，比如宣揚聖誕節戰爭、奧巴馬奪走了美國人手裡的槍等。每次聽到這樣的言論，我就心想，相信這些言論的人該有多傻啊！為什麼要用如此極端的方式談論事情呢？為什麼總是從最壞的角度去想問題呢？但經歷了賈斯汀事件之後，我發現所謂的左派人士也和其他人一樣自以為是，喜歡不加分析地抨擊他人。這個事件的確讓我開了眼界，我最初竟然認為左派人士可能與眾不同，真是大錯特錯。

關於網絡謠言和網絡暴力的科學研究剛剛興起，各種理論尚不成熟。在未來幾年裡，《禿鷲》之類的軟件雖然會繼續完善，但從一定程度上來看，有點類似於《魔域》系列的文字冒險遊戲。不過我一直認為這類研究是非常重要的。我們根據網絡言論分析出來的數據體現了人們內心深處固有的矛盾。人們傾向於對最無力反擊的人發動最猛烈的攻擊。而且最重要的一點是，這些數據揭示了一個長期存在於人們內心深處的願望，即人們希望通過把別人踩在腳下來抬高自己。科學家們已經證明人們這一衝動具有悠久的歷史，但這並不意味著他們對這一衝動有了真正透徹的認識。聖雄甘地曾經說過：「我一直疑惑不解的是：人類怎能通過侮辱自己的同胞而使自己更為榮耀呢？」^[23]

這個問題什麼時候不再令人疑惑呢？讀者也想一想。當人類摒棄了這種惡習，就意味著實現了真正的轉變，因為人類不僅知道了自己具有殘忍的一面，知道了殘忍程度以及什麼時候會表現出殘忍性，也知道了為什麼會存在這種情況。為什麼當一個黑人勝選的時候，人們要去網上搜索關於黑人的笑話呢？為什麼人們喜歡在網絡上發佈一些眼神呆滯、赤身裸體、體態瘦削的人的照片呢？為什麼人們會因為地球的真實年齡而對別人大吼大叫呢？為什麼人們在滿懷深愛的同時又充滿仇恨呢？

[1] 這些文章說薩菲亞的帖子被轉發的次數是1.4萬次，但這些文章是在事情發生後的次日發表的，而我在本書中提到的1.6萬次，則是2014年1月的數據。關於薩菲亞·納瓦茲以及那些給她帶去痛苦的帖子，來源如下：Neetzan Zimmerman, 「Teen Posts Joke on Twitter, Internet Orders Her to Kill Herself,」 Gawker, January 2, 2013, gawker.com/1493156583. Ryan Broderick, 「Meet the

17-Year- Old Girl Who Stood Up to Death Threats After Her Tweet Went Viral on New Year's Eve,」 BuzzFeed, January 2, 2014, buzzfeed.com/ryanhatesthis/meet-the-17-year-old-girl-who-stood-up-to-death-threats-after-ryan-broderick-after-twitter-started-viciously-attacking-her-over-a-silly-joke-this-girl-handled-it-like-a-champ,」 BuzzFeed, January 2, 2014, buzzfeed.com/ryanhatesthis/after-twitter-started-attacking-her-over-a-silly-joke-this-g.

[2] 她們恭賀新年的帖子被轉發的次數是2014年1月的數據，現在肯定被轉發得更多了。

[3] 關於賴格羅開的玩笑以及後續的爭議，我參考了以下資料：「'I'm Not Sorry': Comedian Natasha Leggero Refuses to Apologize Mocking Pearl Harbor Survivors on NBC,」 by that legendary gumshoe 「DAILY MAIL REPORTER,」 Mail Online, January 4, 2014, [dailymail.co.uk/news/article-2533809/Ross-Luippold,Natasha-Leggero's-Stunning-Not-Sorry-Response-over-Controversial-Pearl-Harbor-Joke](http://dailymail.co.uk/news/article-2533809/Ross-Luippold-Natasha-Leggero's-Stunning-Not-Sorry-Response-over-Controversial-Pearl-Harbor-Joke),」 Huffington Post, January 4, 2014, huffingtonpost.com/2014/01/04/natasha-leggero-not-sorry-for-pearl-harbor-joke_n_4541354.html. 發給賴格羅的汗鱗性的帖子，摘自賴格羅本人在Tumblr上發表的一封公開信，這封信的獲取鏈接為：natashaleggero.com/letter/。

[4] 當時，賈斯汀的帖子及其引發的眾怒得到了廣泛報道，有一篇文章就此做了不錯的概述，「Justine Sacco: 5 Fast Facts You Need to Know,」 by Matthew Guariglia, on Heavy, December 21, 2013, heavy.com/news/2013/12/justine-saccoiac-racist-pr-tweet-africa/. 「This Is How a Woman's Offensive Tweet Became the World's Top Story,」 by Alison Vingiano, on BuzzFeed, is a more thorough survey, though one that conveniently omits BuzzFeed's own role in cheering on the mob: buzzfeed.com/alisonvingiano/this-is-how-a-womans-offensive-tweet-became-the-worlds-top-s. 「The Case of Justine Sacco and the Twitter Lynch Mob,」 by Sharon Waxman, in The Wrap, is a piece by someone who, like me, had worked with Justine: thewrap.com/case-justinesacco-twitter-lynch-mob/. 「Justine Sacco: How to Kill a Career with One Tweet,」 by Juana Poareo, is one of many pitiless articles, replete with screenshots of Justine's tweets in the aftermath. The Guardian, 「Liberty Voice」, December 22, 2013, guardianlv.com/2013/12/justine-sacco-how-to-kill-a-career-with-one-tweet/. A screenshot of Google's involvement in #HasJustineLandedYet can be found at 「Justine Sacco Saga Sparks Criticism of Twitter Lynch Mob,」 by Lauren O'Neil, on CBCnews.com: cbc.ca/newsblogs/yourcommunity/2013/12/justine-sacco-saga-sparks-criticism-of-twitter-lynch-mob.html.

[5] 在這一點上，我自己的數據庫裡有成千上萬個充滿惡意的帖子可供選擇，但我還是決定從其他已經得到公開發表的帖子裡面遴選出來幾條。@RonGeraci's tweet appears on his blog, The Minty Plum, in a thoughtful piece, 「View from the Pitchfork Mob,」 January 12, 2014, themintyplum.com/?p=486. @noyokono's tweet appears in Frazier Tharpe, 「PR Woman Tweets Racist Joke Before Flight, Twitter Waits for Her to Land and Get Fired,」 Complex.com, December 21, 2013, complex.com/pop-culture/2013/12/justine-sacco-racist-tweet/. @Kennymack1971's tweet appears in the Sharon Waxman article cited above, 「The Case of Justine Sacco and the Twitter Lynch Mob.」

[6] 阿梅莉亞·埃爾哈特（Amelia Earhart），美國女飛行員，1928年成為第一位獨自飛越大西洋的女飛行員，1937年在飛越太平洋時失蹤。——譯者注

[7] 如果Facebook未來某一天厭倦了那個簡單的f標識，打算換商標，那麼我建議可以採用藍色背景，圖案上是兩個白人在爭論另外一個白人關於非洲的言論。

[8] Alec Hogg, 「Rubbish Rumours.Tweeting Idiot Justine Sacco No Relation to Desmond Sacco, SA Mining Billionaire,」 Biz News.com, December 27, 2013, biznews.com/tweeting-idiot-justinesacco-no-relation-to-desmond-sacco-sa-mining-billionaire/.

[9] 如果研究一下當今世界上仍然採用石刑的國家的人們是否對互聯網版的石刑感興趣，

將是一件非常有趣的事情。

[10] 這項研究並沒有採用我們平時進行隨機化處理過的語料庫；相反，我們選擇了更加完美的數據。為了得到這些數據和有關圖表，我和我的團隊把所有取笑薩菲亞的帖子和涉及「#HasJustineLandedYet」的帖子都遴選了出來。這些數據代表了我對於能夠看到這些帖子的網友的數量所做的最佳推測。

[11] Alan Yu, 「More Than 300 Sharks in Australia Are Now on Twitter,」 All Tech Considered, December 31, 2013, NPR, npr.org/blogs/alltechconsidered/2013/12/31/258670211/.

[12] 他們用信號發射器標記了300多條鯊魚，當這些鯊魚離海岸太近的時候，發射器就會自動發出一條Twitter來警告鯊魚威脅。這個用於警告鯊魚威脅的Twitter賬號叫作@SLSWA，為晒日光浴和海上衝浪的遊客提供鯊魚行蹤的相關信息。

[13] My source for the history and science of rumors is Jesse Singal's piece 「How to Fight a Rumor,」 Boston Globe, October 12, 2008, boston.com/bostonglobe/ideas/articles/2008/10/12/how_to_fight_a_rumor/.他提出的一個深刻見解就是將謠言與網絡惡意聯繫在一起。他也引用了《聖經》裡面「藐視鄰舍的，毫無智慧……」這一句話。不過這一段接下來的「若不想被人評判，自己最好不要評判別人」和「謠言女神」則是我自己引用的。I also used 「Rumor, Gossip and Urban Legends,」 by Nicholas DiFonzo and Prashant Bordia, in *Diogenes* 54, no.1 (2007): 19—35, and Mr.DiFon-zo's article 「Rumour Research Can Douse Digital Wildfires」 in *Nature* 493, no.7431 (2013): 135.

[14] 最初是在線漫畫網站「娛樂場」（Penny Arcade）引導我關注蘇勒的工作。關於蘇勒和「網絡抑制解除效應」的一些基本觀點，我參考了維基百科上面「Online disinhibition effect」條目的內容。這個條目也提到了「娛樂場」這個在線漫畫網站，該網站的網址為：pennyarcade.com/comic/2004/03/19。

[15] 當時那些卡車司機甚至也有自己的別名，就像今天的網友給自己取的網名一樣。

[16] 關於這一點，我參考的數據來源是維基百科上的「Online disinhibition effect」條目。條目內容引用的原始數據來源是Kenneth Tynan於1978年2月20日在《紐約客》雜誌上發表的「Fifteen Years of the Salto Mortale」一文。

[17] 如果想進一步瞭解與打電話有關的幽默，我為您推薦Longmont Potion Castle這部電視劇，裡面有很多利用電話完成的惡作劇。

[18] 「阿拉伯之春」可謂Twitter在社會運動中的首秀，而Twitter則變成了一個具有全球重要性的工具。在危地馬拉、摩爾多瓦、俄羅斯和烏克蘭的抗議運動中，Twitter也起到了推動作用。

[19] See Todd Dugdale, 「Sandbaggers and Trolls,」 *kd0tls Ham Radio Experience*, January 6, 2014, kd0tls.blogspot.com/2014/01/sandbaggers-and-trolls.html/.

[20] 這些數據來自我隨機處理過的數據樣本。

[21] 在分析Facebook的數據時，使用的數據都是來自經過隨機和匿名化處理的數據庫。與「任何人都不應……」有關的討論和列表，引自拉達·阿卡米克（Lada Adamic）等人於2014年1月18日發表的「The Evolution of Memes on Facebook」一文，該文獲取鏈接為：facebook.com/notes/facebook-data-science/the-evolution-of-memes-on-facebook/10151988334203859。這篇文章並沒有明確說明政治偏見的判斷標準，我的推測依據是用戶的點贊情況。

[22] 本段討論美國政治的極端化，請參考吉爾·萊波雷（Jill Lepore）於2013年12月2日在《紐約客》上發表的「Long Division」一文。

[23] 自從2007年讀了路易斯·費舍爾（Louis Fisher）為甘地撰寫的傳記《甘地傳》（Life

of Mahatma Gandhi) (New York: Harper & Brothers, 1950) 之後，甘地這句話就一直縈繞在我的腦海裡。

第三部分 影響身份認同的因素

第十章 你是誰？

在申請大學的時候，我必須寫一份自我陳述。我現在已經不記得當時申請表上的問題了，其實，無論問題是什麼，其實都無關緊要。它的目標就是讓我描述一下自己的情況，這樣負責招生工作的人可以判斷一下他們是否喜歡自己看到的內容，就像現在的通用申請表所說的那樣：「你的個人陳述有助於我們瞭解你。」

當時，我非常喜歡情景劇。於是，我在自我陳述中寫到每當我去上學而不得不與我的小狗分別時，就會非常傷心。那個小狗叫「霜霜」（Frosty），我6歲時就得到它了。因此，可以說它是和我一起成長的，但它老得太快了。我們家搬了好幾次，俱樂部、社區游泳池以及小朋友都無法跟我一起走，都留在了休斯敦、克利夫蘭或路易斯維爾，但無論我們搬到哪兒，「霜霜」一直跟著我。因此，它是我對童年最後的眷戀，但我知道自己不得不繼續獨自前行了。

當時，我渾身散發著憂鬱氣質，喜歡穿超大號的T恤衫，上面印著荷蘭藝術家埃舍爾（M.C. Escher）的魔法圖形。就這樣，我完成了我的大學申請。自那之後，我就沒怎麼寫過個人陳述了。因為我現在從事社交網站的管理工作，能看到無數用戶寫的自我陳述，所以，經常忍不住回想起17歲時的自己，以及當年自己寫的那份自我陳述。我不禁會想，自己當時感興趣的事情那麼多，為什麼只談到那隻小狗呢？為什麼不談論棒球、籃球、網球呢？當那張申請表上提示「你是誰」時，究竟是什麼讓我做出了那樣的回答呢？更加重要的是，其他孩子當時是如何回答這個問題的呢？

20多年後的今天，作為社交網站的管理者，我能接觸到數千萬名用戶寫的自我陳述，總字數多達數十億。這些陳述大體上回答了同樣一個問題，這個問題就是：「你是誰？」在申請大學時，我是一名申請者，我寫的個人陳述是給大學招生人員讀的；現在，我的角色發生了轉變，

我不再是一名申請者，而是網站管理者，能夠讀到他人的自我陳述。但那些大學招生人員在讀申請者的陳述時，會拿一個預先設定好的理想標準去評判別人，看看申請者是不是一塊兒讀大學的好料，但我卻沒有拿固定的標準去衡量他人。我可以把這些自我陳述彙集到一起，它們能為我揭示出作者的理想是什麼。有時候，一組數據非常詳細，你不需要提出任何問題，這些數據就能給你透露出作者的一切。人們是如何描述自己的？哪些事情是重要的？哪些是典型的？哪些是非典型的？如果每個人都有機會寫一下自己究竟是誰，那麼他們會著重描繪自己的哪些特徵？

在這裡，我們將著重分析幾大類人，比如黑人、白人、亞裔、女性、男性等。在研究某一類人的時候，一個問題就是你總會帶著自己的偏見和先入之見。你看到的、記住的和記錄下來的東西既取決於現實，也取決於你的分析視角。在盛水時，你用什麼形狀的桶，水就會呈現出什麼樣的形狀。在社會科學領域，知識就像水一樣，你用什麼樣的視角去分析，就會得出什麼樣的結論。因此，如果我們要通過我收集的這些自我陳述去客觀地瞭解作者們的信息，比如關於其種族、性別和性取向的信息，那麼我們就需要開發出一種數學算法，去除我們的主觀視角，用客觀視角來研究他們。

如果讀一讀OkCupid網站的用戶提交的自我陳述，你就會發現它們非常類似於個人總結。網站只給他們提供了一些非常簡短的和開放式的提示，包括：

「我的自我總結.....」

「我非常擅長.....」

「我最吸引人的地方是.....」

「我花很多時間思考.....」

用戶往往在自我陳述中努力展現出最好的一面，從這一點來講，與申請大學時的自我陳述沒什麼區別。我想很多人在寫這類東西的時候，總會感覺有些焦慮，因為網站除了簡短的提示之外，並沒有設定長度限制，也沒有提供指導原則。總體來看，所有用戶在該網站上留下了多達32億單詞的自我陳述。此外，與谷歌圖書等浩如煙海的數據庫相比，我收集的這些文本中的每個單詞背後都反映了作者的一些人口學信息，比如年齡、居住地點、種族等。在之前的研究中，我們總是傾向於統計哪個詞語出現的頻率最高，即詞頻最高。然而，如果要從一個群體（比

如，亞裔女性群體）的文本中總結出一個能夠獲得該群體認同的特徵，卻並非如此簡單。如果我們只是統計詞頻的話，就會得出下面的結論：定冠詞the的詞頻排名第一，介詞of的詞頻排名第二，連詞and的詞頻排名第三.....這樣排下去的結果與牛津英語語料庫中詞頻排在前100名的單詞沒什麼區別。在所有講英語的群體中，無論是亞裔女性，還是白人男性，抑或其他群體，在表達自己的意思時，使用的代詞、冠詞、介詞是相同的。因此，如果僅僅統計詞頻，這些居於基礎地位的單詞肯定排在前列。所以，要發現某一個群體的獨特之處，也就是說只有他們這一個群體才有而其他群體沒有的因素，我們就需要先用一種不同的方法對文本進行分類。^[1]

我姑且使用「白人男性」這個群體為例吧，因為我對這個群體瞭解得最清楚。在開始分析之前，第一步是將「白人男性」用戶寫的自我陳述與其他人的自我陳述區分開，也就是說，將所有的自我陳述分為「白人男性」與「其他人」。然後，我們根據這兩類文本中的詞和短語出現的頻率按照從高到低進行排序，就得到了圖10—1。這兩類文本各有36萬個詞和短語，我遴選出the（這個、那個）、pizza（比薩）、Phish（費西合唱團）這三個詞作為例子，根據詞頻將其準確地排列到了相應位置。

白人男性的自我陈述

其他人的自我陈述

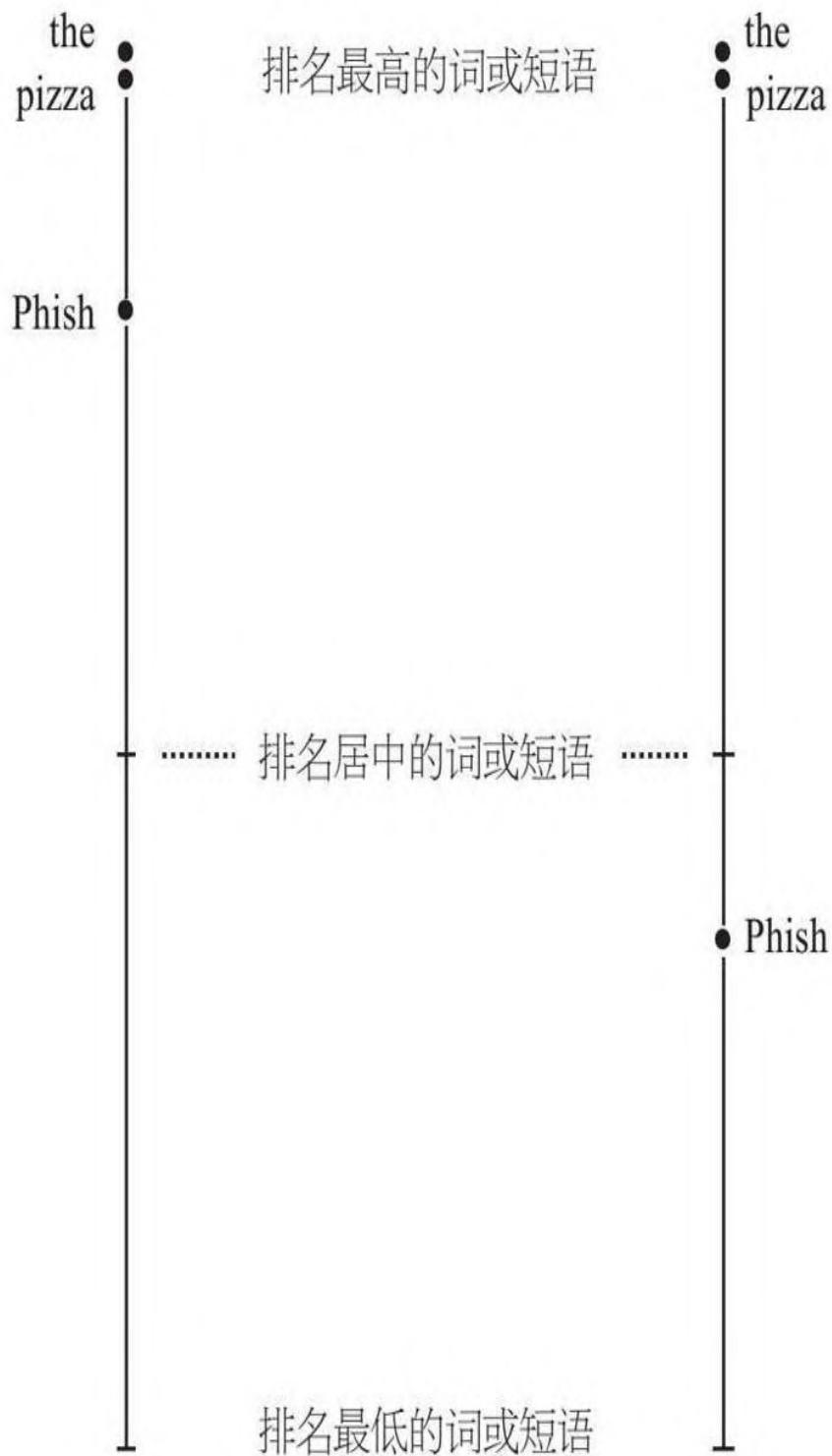


圖10—1 白人男性與其他人自我陳述中高頻詞對比

這樣，我們的分析已經開始有了一些進展，但在深入開展分析之前，我想先趁圖10—1看起來比較簡單時解釋一下一個可能會引起誤解的地方。雖然Phish這個詞的位置讓很多人困惑不解，但我要說的與它無關。我所擔心的是，pizza的詞頻竟然與the的詞頻幾乎處於同一水平，這一點令人費解，很容易讓人產生誤解。按理講，the是個定冠詞，高居榜首肯定不會引起疑慮，而pizza只是一種頗受歡迎的食物，但在我們的數據中，二者的百分位數都是98，也就是說，有98%的詞或短語出現的頻率等於或低於the和pizza出現的頻率。這樣，人們可能會以為我的數據存在疏漏或者我採用的統計方法存在疏漏，但事實上，這樣的排名結果是正確的。之所以出現這樣的現象，是因為人們在使用語言的過程中存在一個奇怪的傾向，即我們總是在重複使用同樣的詞來表達自己的意思，這就導致一些高頻詞在我們的文章裡佔據了絕大部分。這樣，除了那些詞頻排名最高的詞之外，詞頻排名稍低的詞的使用次數便驟然少了很多。

一個詞的受歡迎程度（即其在語料庫中的出現次數排名）與其出現次數之間的關係似乎違反人們的直覺。齊普夫定律（Zipf's law）描述了這種關係。所謂齊普夫定律，又稱為「單詞分佈定律」，在對大量文獻進行文本分析的基礎上，把文獻中的詞按照詞頻由高到低排列，發現在自然語言的語料庫裡，出現次數與其詞頻排名成反比，二者之乘積近似於一個常數。^[2]這條定律是在觀察和分析大量文獻的基礎上總結出來的一個語言領域的統計學特徵，如同其他基於經驗數據的定律一樣，看似巧合的外表下蘊含著某種奇蹟般的規律。^[3]

這條定律適用於《聖經》，適用於20世紀60年代的流行歌曲歌詞集，也適用於牛津英語語料庫這一經典的英語數據庫，當然也適用於我收集的海量自我陳述文本。接下來，我們以詹姆斯·喬伊斯（James Joyce）的《尤利西斯》這個風格獨特的文本為例，真切地看一看這條定律的適用情況（見表10—1）。^[4]

表10—1 《尤利西斯》的詞頻排名

单词	词频排名	出现次数	词频排名与出现次数之积
's (is、was、us、has等的缩写)	10	2 826	28 260
is (是)	20	1 435	28 700
what (什么)	30	975	29 250
has (拥有)	100	289	28 900
wife (妻子)	200	140	28 000
Ireland (爱尔兰)	300	90	27 000
college (大学)	1000	26	26 000
morn (早上)	5000	5	25 000
builder (建筑师)	10000	2	20 000
Zurich (苏黎世)	29055	1	29 055

詞頻排名與出現次數之間的關係大體上保持穩定，這似乎既是一個語言特徵，又是一個人類思維的特徵，因為你在表10—1中可以看到既包含「愛爾蘭」「蘇黎世」這樣的專有名詞，也包含像「's」這種從個人習慣用語中摘出來的構詞元素。

這條定律還可以用於描述多種多樣的社會架構，包括城市規模和收入分配情況。以城市規模為例，世界上很多國家的數據都證明一國最大

的城市的人口，是第二大城市人口的兩倍，是第三大城市人口的三倍，依此類推，即最大城市的人口是第N大城市的N倍。這進一步印證了該定律與人類思維之間存在著深刻聯繫。就我們所討論的內容而言，這條定律給我們的啟示是，在大多數語料庫中，少量高頻詞反覆出現，而其他詞的出現次數則迅速降低，可以說20%的詞佔了80%的出現次數。換言之，一個詞使用的次數越多，則其再次被使用的可能性就越大。在我收集的個人陳述中，幾乎每篇陳述都會用到the這個單詞，或者說the一詞的詞頻大約是100%。相比之下，pizza一詞的詞頻大約是1/14。雖然對於白人男性而言，Phish一詞的詞頻處於第80個百分位，或者說在白人男性的用語中，Phish一詞的詞頻高出了80%的詞，但其詞頻還不到1/200。現在，分析了詞頻排名和出現次數之間的關係之後，接下來我們就要在這種關係的基礎上，朝著我們的目標展開更加深入的分析。

我把圖10—1中的縱軸與橫軸交叉在一起，形成了一個直角，呈現出一個正方形的形狀，就得到了圖10—2。在圖10—2中，縱軸表示一個詞在白人男性群體中的受歡迎程度，橫軸表示一個詞在其他人中間的受歡迎程度。我在Phish這個詞的周圍添加了兩個箭頭，這樣我能更加清楚地表達自己的意思。

一個詞在圖10—2中的位置具有雙重含義，越靠上表示其越受白人男性的歡迎，越靠右表示其越受其他人的歡迎。在擴大到所有詞之前，我先在圖10—2中稍微添加幾個單詞，這樣就有助於你瞭解到這個幾何圖形的變化情況（見圖10—3）。

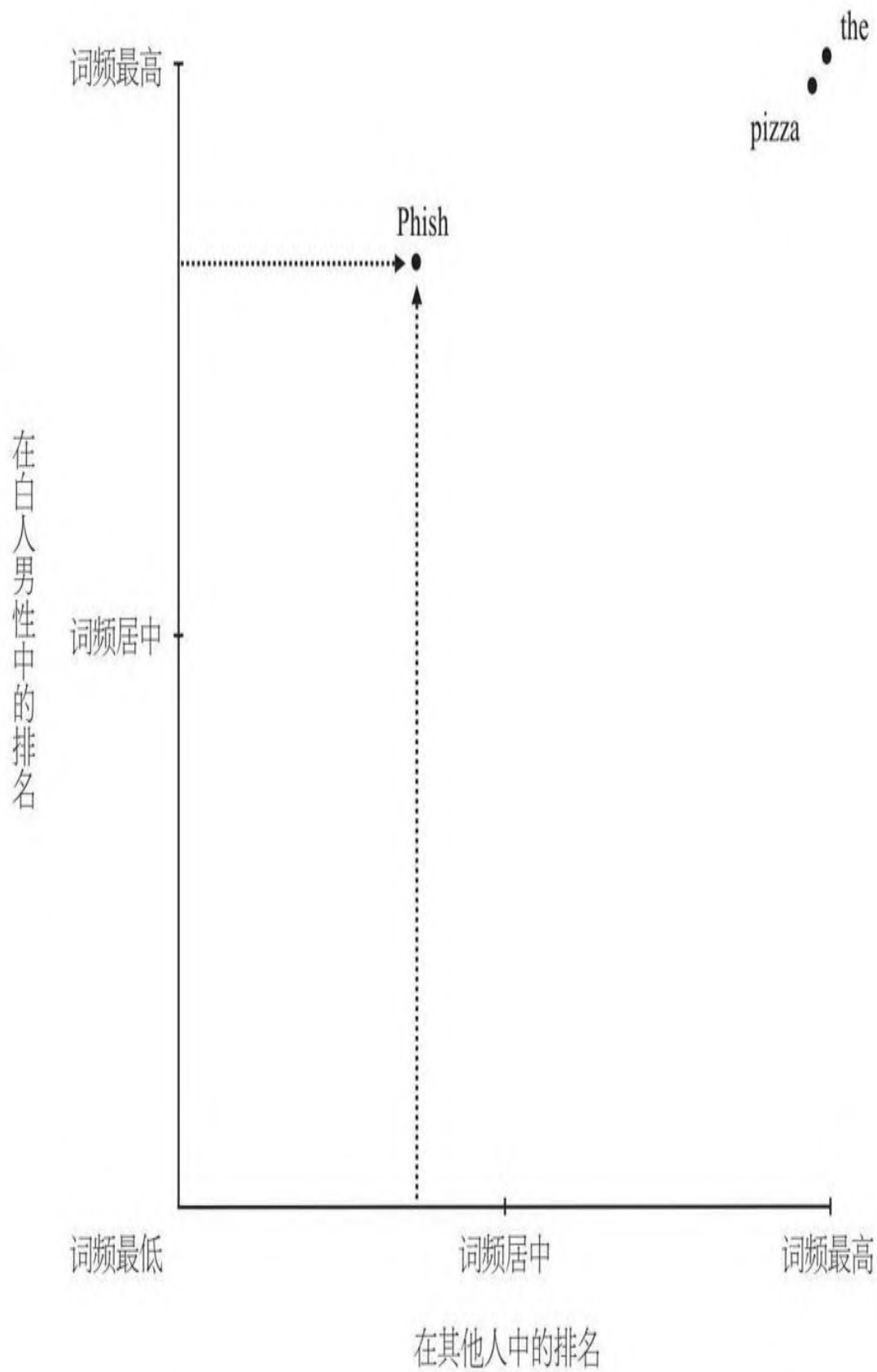


圖10—2 白人男性與其他人自我陳述中高頻詞對比變形圖

我在圖10—3中還添加了一條對角線，一個詞越靠近對角線，表示其在白人男性和其他人群中的受歡迎程度越接近；一個詞的位置越靠近右上方，表示其越具有普遍重要性。但請記住，我們不是在尋找兩類人群之間的共性，而是在尋找差異性。我們在這裡進行分析的最終目標是瞭解白人男性群體具有哪些特殊之處。為了達到這個目的，我們需要將視線停留在圖10—3的左上角：一個詞越靠近左上方，白人男性群體使用它的頻率越高，而其他人群使用的頻率則越低。事實上，如果一個詞位於左上角，也就是位於正方形左上方的頂點上，那麼就表示這個詞是白人男性群體的專屬，只有白人男性群體使用它，而其他人群則不會使用。請想象一個位於左上方頂點的詞，每個白人男性的自我陳述中都會出現這個詞，而其他人的自我陳述中則從來不會出現它。如果這個詞是在自我總結中出現的，那麼用這個詞來界定一個群體的集體特徵是非常理想的。根據這個方法，一個詞距離左上方頂點的距離就像能夠開口說話一樣，能夠幫助我們瞭解一個群體是如何談論自己的。

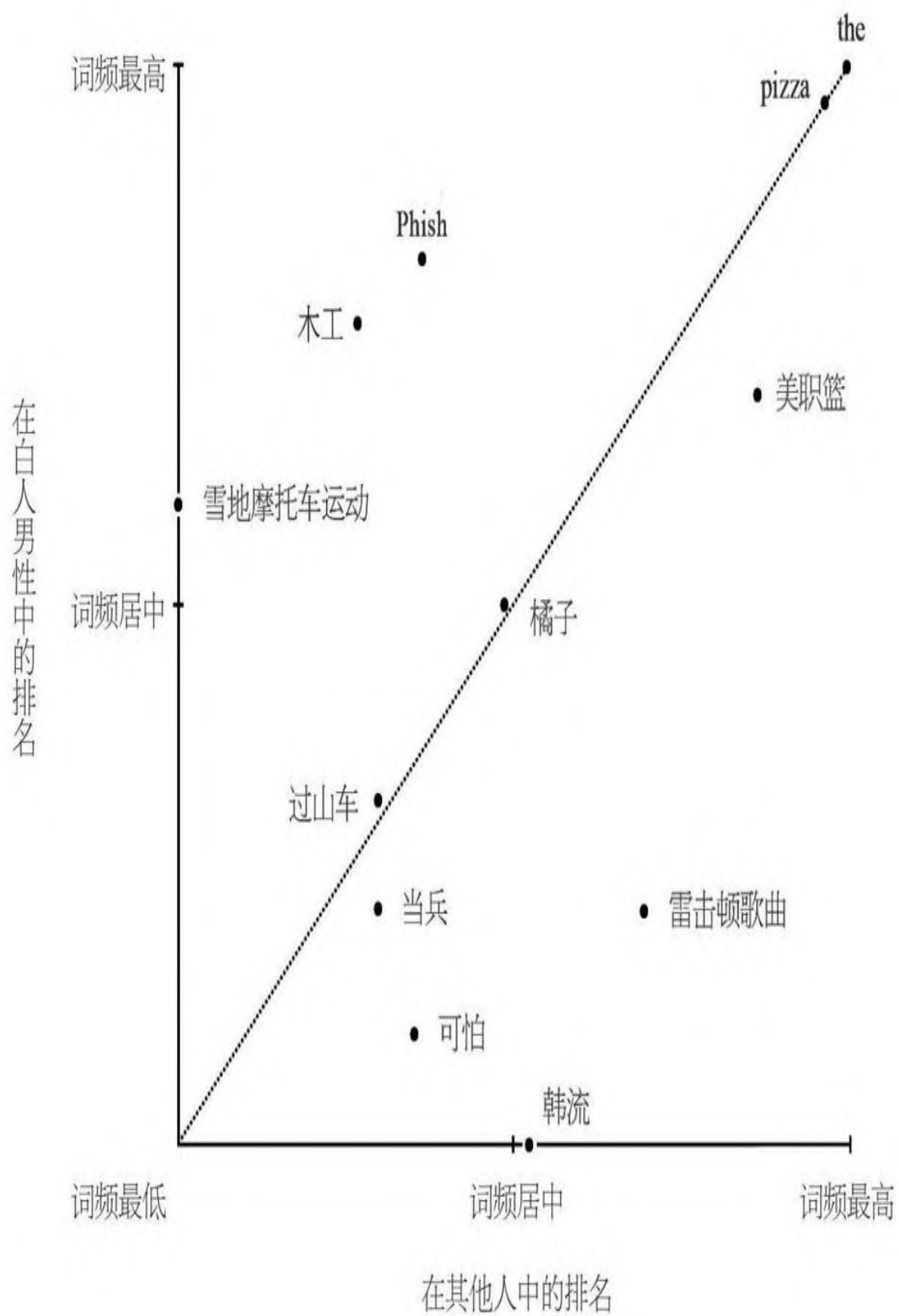


圖10—3 白人男性與其他人自我陳述中高頻詞對比（詞條擴大後）

但每一組數據都有其特殊之處，因此，研究人員必須經常重新建立分析工具。無論任何時候，如果你建立了一個分析工具，那麼你最好用一些自己熟悉的現象來檢驗一下分析結果是否符合現實。如果符合現實，則表明你的方法是正確的；反之，則分析方法不正確。這就像一個木匠造了一條新船，在駛入茫茫大海之前，誰也不知道會發生什麼，所以最好在岸邊時先檢查一下是否存在漏洞。我建立了這個分析工具之後，如果通過這個工具發現「韓流」或「可怕」這兩個詞位於左上方，那麼根據我對白人男性的瞭解，這就有力地表明要麼我的數據出錯了，要麼我的分析工具出錯了。但正如你們能看到的那樣，一切分析結果都非常符合白人男性的現實狀況。

因此，最後把我收集的這兩個群體自我陳述中出現的所有詞都放到這個圖中，就得到了下面的結果（見圖10—4）：

在圖10—4中，距離左上角最近的那個圓點代表的是「我的藍眼睛」（my blue eyes），這是最具白人男性特徵的一個詞語。如果我們想尋找更多專屬於白人男性的詞，那麼就從左上方的頂點向外看，比如，距離這個頂點最近的前30個圓點就代表白人男性最常用的30個詞。這樣一來，我們就能利用幾何圖形來找到白人男性群體中的常用詞。

我不僅僅為白人男性群體繪製了這樣的圖形，我還為數據中的其他群體繪製了相應的圖形，而且正是通過同樣的數學方法，我還找到了各個群體的轉述詞。但在列出所有這些詞之前，我想先說明非常重要的一點。如果按照性別（2）、種族（4）和性取向（3）來分析，那麼你就能把所有人分為24類，即 $2 \times 4 \times 3 = 24$ 。相應地，你就會繪製出24個類似的圖形，而且在所有圖形中，中間的重疊部分總是從左下方向右上方不斷變窄。也就是說，一個單詞的位置越靠近右上方，那麼它越接近對角線，這就意味著，我們傾向於在最重要的事物上達成共識。至於我們無法達成共識的事物，我在表10—2中詳細地列了出來。我先從「男性」開始。

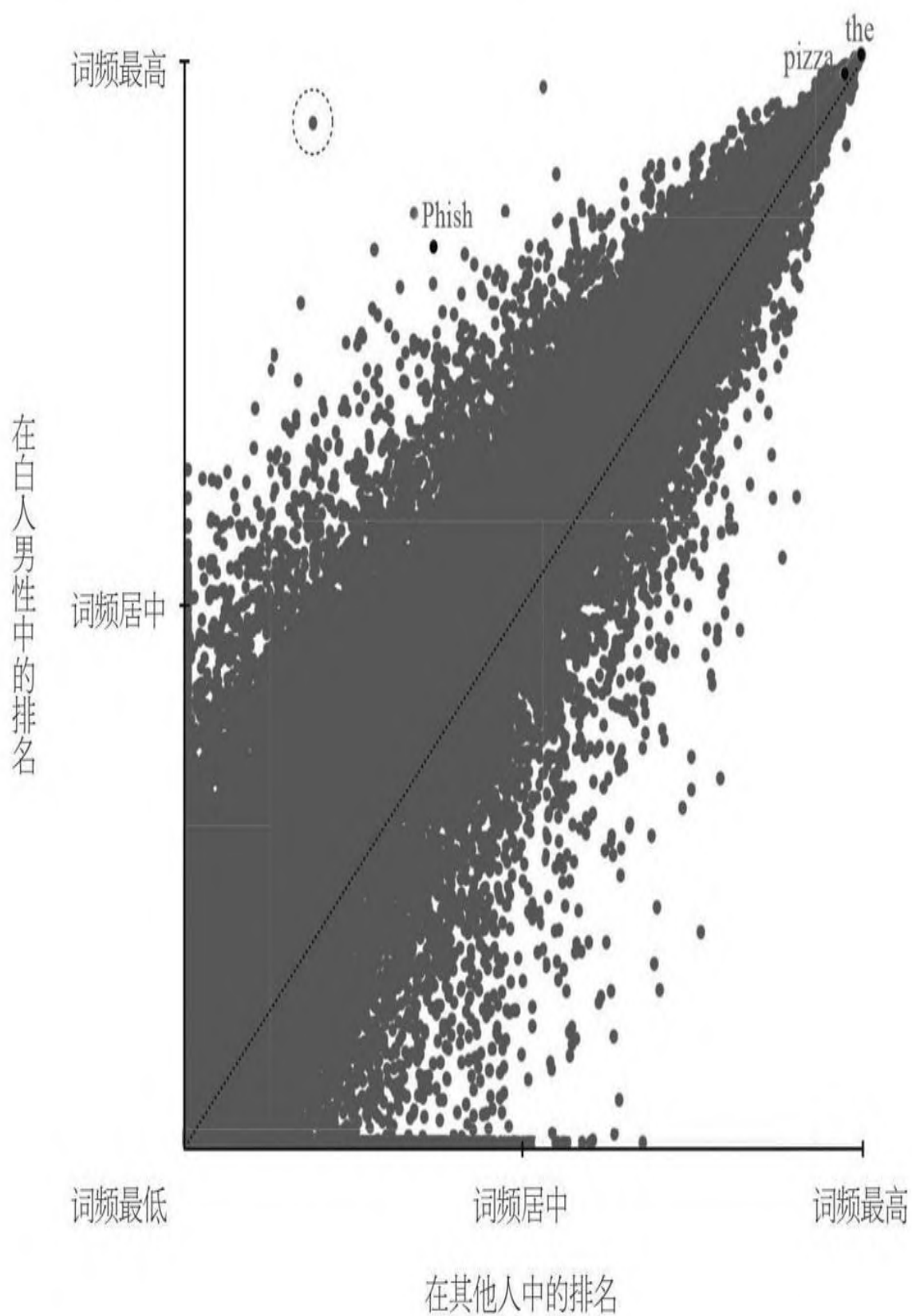


圖10—4 白人男性與其他人自我陳述中高頻詞對比（所有詞）

表10—2 不同族群男性使用最多的詞

白人男性	黑人男性	拉丁裔男性	亚裔男性
my blue eyes (我的蓝眼睛)	dreads (害怕)	colombian (哥伦比亚人)	tall for an asian (对亚裔而言高了些)
blonde hair (金发)	jill scott (吉尔·斯科特)	salsa meringue (莎莎梅伦格舞)	asian (亚裔)
brown hair (棕色头发)	haitian (海地人)	cumbia (空比亚舞)	cantonese (广东人)
hunting and fishing (打猎和捕鱼)	soca (索卡音乐)	una (唯一的)	infernal affairs (无间道)
allman brothers (阿尔曼兄弟乐队)	neo soul (新灵魂乐)	merenguebachata (梅伦格巴恰塔舞)	seoul (首尔)
woodworking (木工活)	jamiefoxx (杰米·福克斯)	mana (魔力)	infernal (地狱般的)
campfire (野营)	zane (上帝恩典)	banda (班达舞)	shanghai (上海)
dropkick murphys (普考斯·墨菲斯乐队)	paid in full (电影《全款交收》)	puertorican (波多黎各人)	boba (波霸)
they might be giants (明日巨星合唱团)	nigga (黑人)	colombia (哥伦比亚)	kbbq (康伯巴奇)
brewing beer (酿制啤酒)	luther vandross (路德·范德鲁斯)	gusta (喜欢)	kpop (韩流)
robert heinlein (罗伯特·海因莱因)	coldest winter (图书《最寒冷的冬天》)	corridos (民谣)	badminton (羽毛球)
tom robbins (汤姆·罗宾斯)	tyler perry (泰勒·派瑞)	bachatamerengue (巴恰塔梅伦格舞)	kimchi (泡菜)
townes (歌星汤尼)	swagg (斯瓦格)	hector (威吓)	chungking express (电影《重庆森林》)
old crow medicine show (“老鸦医药秀”乐团)	jerome (杰罗姆)	espa (西班牙)	chou (周)
mystery science theater (神秘科学剧场)	dreadlocks (细发辫)	por (葡萄牙)	viet (越南)
skis (滑雪板)	spike lee (斯派克·李)	salsa bachata (莎莎巴恰塔舞)	jiro (次郎)

(續表)

白人男性	黑人男性	拉丁裔男性	亚裔男性
sailboat (帆船)	holla at me (电影《罪无可赦》)	aventura (多米尼加乐队)	dash berlin (达许·柏林乐队)
around a fire (围着篝火)	menace to society (电影《社会威胁》)	english and spanish (英语和西班牙语)	ucsd (加州大学圣迭戈分校)
caddyshack (电影《凤凰高尔夫》)	brotha (兄弟)	musica (音乐)	beijing (北京)
blond hair (金发)	shottas (牙买加黑帮)	espa ol (西班牙语)	hk (香港)
bill bryson (比尔·布莱森)	nigerian (尼日利亚的)	como (因为)	norwegian wood (小说《挪威的森林》)
wheelers (精明的人)	heartbeats (心跳)	fiu (佛罗里达国际大学)	jiro dreams of sushi (电影《次郎的寿司梦》)
pogues (棒克乐团)	anthony hamilton (安东尼·汉密尔顿)	pero (秘鲁)	lin (林)
barenaked ladies (裸体淑女乐队)	gud (上帝)	soledad (孤独)	philippines (菲律宾)
mst3k (美剧《神秘科学剧院》)	wayans (韦恩斯)	espanol (西班牙人)	noodle soup (面汤)
truckers (艾斯里兄弟合唱团)	dickey (迪基)	amor (爱情)	malaysian (马来西亚人)
jethrotull (杰思罗·塔尔)	interracial (混血的)	muy (非常)	for my next meal (为了我的下一顿饭)
canoe (划独木舟)	nigeria (尼日利亚)	reggaeton (雷击顿音乐)	gangnam style (歌曲《江南风》)

白人男性的常用詞中，有很多與音樂有關，比如「費西合唱團」等，這揭示出在白人男性的內心深處湧動著對音樂的熱愛。其他三類男性的常用詞列表中，很多都是我之前從來沒有聽說過的，比如「上帝恩典」「安東尼·漢密爾頓」「《最寒冷的冬天》」「《重慶森林》」「達許·柏林樂隊」等。正是得益於我的數學分析方法，才把這些詞揭示出來。事實上，花上幾分鐘時間，到維基百科上也能查到一些這類詞，但我並不認為這樣就能全面清楚地瞭解一種文化。我在表10—2中列出的這些詞都是這些人群在日常生活中使用的，在此基礎上，我們可以看出下面這個大趨勢：白人主要傾向於通過頭髮和眼睛來描述自己，將自己與他人區別開來，亞裔傾向於通過自己原本所處的國家或地區來區別自己，拉丁裔則傾向於通過自己的音樂來區別自己。作為一名白人男性，我對其他三類人的文化不甚瞭解。雖然我熟悉「斯派克·李」「北京」「上海」等少數幾個詞，但有了這個列表，我就可以從一個「內部人士」的視角去觀察和分析他們的文化了。如果沒有這種分析方法，一個外人是無法通過搜索引擎的自動完成功能去得到這些數據的，因為你根本就不知道這些詞的存在，也不知道要搜索什麼。對於日本作家村上春樹的《挪威的森林》一書以及根據此書改編的電影，如果不是亞裔，可能絕大多數人都不熟悉。如果讓我猜，我可能會猜測這是甲殼蟲樂隊的一首歌。如果有人在我寫這一章之前問我是否看過這部電影，我可能會回答說：「生活在森林裡的挪威人會拍電影嗎？」但通過建立的這個數學分析法有了上面這個列表，我們就能明白亞裔對於這本書和這部電影的熱情了。我之前並不知道亞裔對這本書和這部電影的熱情，而是有了這些數據之後才認識到的。這些詞都是不同文化的人特有的，因此，其他文化的人不可能憑空造出來，即便利用「谷歌趨勢」或搜索數以百萬計的標籤，也不會發現這些詞的存在。事實上，我們要發現這些數據，有時需要採用我這種「盲算法」（blind algorithm）。

表10—3是4類「女性」的常用詞。根據這個列表，你能看到這些女性在總體精神風貌上類似於同一個文化背景下的男性，只是多了一些與抒情歌曲有關的詞。

表10—3 不同族群女性使用最多的字詞

白人女性	黑人女性	亚裔女性	拉丁裔女性
my blue eyes (我的蓝眼睛)	soca (索卡音乐)	tall for an asian (就 亚裔而言算高的了)	latina (拉丁裔)
red hair and (红头发)	eric jerome dickey (埃里克·杰罗姆·迪基)	philippines (菲律宾)	colombian (哥伦比亚人)
blonde hair and (金发)	haitian (海地人)	beijing (北京)	una (唯一)
love to be outside (喜欢户外)	imitation of life (电影《生活的模仿》)	coz (表亲)	cumbia (空比亚舞)
mudding (渣滓)	zane (上帝恩典)	boba (波霸)	banda (班达舞)
campfie (篝火)	coldest winter ever (小说《最寒冷的冬天》)	fiipina (菲律宾人)	tejano (特哈诺音乐)
four wheeling (开车兜风)	nigerian (尼日利亚人)	cantonese (广东人)	merenguebachata (梅伦格巴恰塔舞)
phish (费西合唱团)	interracial (混血的)	asians (亚洲人)	gusta (喜欢)
hunting fishing (打猎和捕鱼)	rb and gospel (节奏布鲁斯与福音音乐)	wong kar wai (王家卫)	puertorican (波多黎各人)
campfires (篝火)	five heartbeats (电影《无心合唱团》)	shanghai (上海)	colombia (哥伦比亚)
green eyes and (绿色的眼睛)	anita baker (安妮塔·贝克)	seoul (首尔)	mana (魔力)
auburn (红棕色)	brooklyn (电影《种族情深》)	macarons (法式小圆饼)	vida (生活)

(續表)

白人女性	黑人女性	亚裔女性	拉丁裔女性
ride horses (骑马)	neosoul (新灵魂乐)	viet (越南人)	bachata merengue (巴恰塔梅伦格舞)
grateful dead (感恩而死乐队)	octavia butler (奥克塔维亚·巴特勒)	kimchi (泡菜)	amor (爱情)
mountain goats (山羊)	housewives of atlanta (真人秀《亚特兰大贵妇的真实生活》)	for my next meal (为了我的下一顿饭)	musica (音乐)
love country music but (热爱乡村音乐)	luther vandross (路德·范德鲁斯)	singapore (新加坡)	english and spanish (英语和西班牙语)
gillian welch (吉兰·威尔奇)	zora (卓拉)	malaysian (马来西亚人)	español (西班牙人)
country girl (乡村女孩)	waiting to exhale (电影《待到梦醒时分》)	hk (香港)	salsa merengue (莎莎梅伦格舞)
christmas vacation (圣诞假期)	anthony hamilton (安东尼·汉密尔顿)	malaysia (马来西亚)	todo (待办事项)
bill bryson (比尔·布莱森)	chrisette (克利赛特)	noodle soup (面汤)	por (葡萄牙)
riding horses (骑马)	locs (黑人卷发)	cambodian (柬埔寨人)	mariachi (墨西哥流浪乐队)
eric church (埃里克·丘奇)	outside my race (其他种族)	norwegian wood (小说《挪威的森林》)	marc anthony (马克·安东尼)
barn (车库)	real housewives of atlanta (真人秀《亚特兰大贵妇的真实生活》)	hong kong (香港)	español (西班牙语)
allman (阿尔曼兄弟乐队)	calypso (卡吕普索舞)	chungking express (电影《重庆森林》)	novelas (小说)
willie nelson (威利·纳尔逊)	know why the caged (图书《我知道笼中鸟为何歌唱》)	rachmaninoff (拉赫玛尼诺夫)	venezuela (委内瑞拉)
harley (哈雷)	did i get married (电影《后悔莫及》)	southeast asia (东南亚)	soledad (孤单)
brunette (褐色头发)	spike lee (斯派克·李)	vienna (维也纳)	mas (化装舞会)
flogging molly (弗根莫利乐队)	braxton (布莱克斯顿)	mandarin (中国普通话)	tacuba (塔古巴乐团)

在分析過程中，我發現我採用的這套數學分析法非常靈活，可以逆向使用，也就是說，既可以用它來統計出一個群體最喜歡用哪些詞，也可以統計出一個群體最不喜歡用哪些詞。這樣的統計結果可以讓你發現一個群體的另一面，同樣具有啟發意義。接下來，我列出了男性最不喜歡使用的4類詞（見表10—4）。背景顏色之所以設置為灰色，是為了在視覺上突出這裡列出的內容與前面列出的內容截然相反。這些詞都是某個群體最不常用而其他群體經常使用的。根據這些內容，你可以判斷出各個群體總體上的精神風貌。這些列表值得仔細研讀。

表10—4 不同族群男性使用最少的詞

白人男性	黑人男性	亚裔男性	拉丁裔男性
slow jams (慢拍舞步)	borges (博尔格斯)	sence (感觉)	southern accent (南方口音)
trey songz (黑人说唱歌手崔·尚斯)	social distortion (畸世乐团)	layed (打赌)	from the midwest (来自中西部)
robin thicke (罗宾·西克)	tallest man on earth (地球上最高的人)	layed back (优哉游哉)	ann arbor (安阿堡市)
smh (州立精神病院)	gaslight anthem (煤气灯圣歌乐队)	sence of humor (幽默感)	midwestern (中西部)
musiq (音乐)	snorkeling (浅滩潜水)	truck driver (卡车司机)	gumbo (冈波语)
laker (湖畔派诗人)	xkcd (网络漫画名称)	realy (真正地)	equity (公平)
ig (不理睬)	diet coke (健怡可乐)	anything else you wanna (你想要别的吗)	discworld (小说《荒诞世界》)
kevin hart (凯文·哈特)	surfboard (冲浪板)	like what u see (就像你看到的那样)	shanghai (上海)
raised in nyc (在纽约长大)	totoro (龙猫)	and my son (我儿子)	scallops (扇贝)

(續表)

白人男性	黑人男性	亚裔男性	拉丁裔男性
hip hop rap rb (嘻哈说唱)	magnetic filds (磁场)	u like what u (你喜欢你……的)	slopes (溜掉)
kpop (韩流)	gogol bordello (果戈理妓院乐队)	care of my kids (照顾我的孩子)	university of michigan (密歇根大学)
george lopez (乔治·洛佩兹)	dropkick murphys (普考斯·墨菲斯乐队)	making (制造)	assessment (评估)
neo soul (新灵魂乐)	rebelution (游戏《嘻哈狂潮》)	welder (焊工)	parentheses (括号)
rb and hip hop (嘻哈说唱)	peru (秘鲁)	hunting fihing (打猎与捕鱼)	snowboarder (滑雪板)
neyo (黑人歌手尼欧)	dr horrible's sing along blog (恐怖博士的欢唱博客)	care of my son (照顾我的儿子)	nyt (《纽约时报》)
knw (知道)	wakeboarding (水上滑板)	wanna know anything else (想了解其他 事情)	dominion (统治)
gud (上帝)	herzog (赫索格)	else you wanna know (想知道其他事情)	msu (密歇根州立 大学)
follow me (跟我来)	my blue eyes (我的蓝眼睛)	raising my son (养育我的儿子)	ellipses (椭圆)
jordans (夜壶)	guitar and sing (吉他弹唱)	ask and ill (提问, 我将……)	maple (枫叶)
handball (手球运动)	dr horrible's sing along (恐怖博士的欢唱博客)	comedys (喜剧)	nigerian (尼日利亚人)
soulchild (顽童)	coachella (柯契拉音乐节)	dnt (禁止跟踪)	kenya (肯尼亚)
ne yo (黑人歌手尼欧)	dr horrible's sing (恐怖博士的欢唱博客)	woman who wants (想要……的女子)	john irving (约翰·欧文)
bachata (巴恰塔舞)	yo la tengo (优拉糖果乐团)	i'm a single father (我是一名单身父亲)	over a decade (在十年时间里)
basketball (篮球)	airborne toxic event (毒害漫延合唱团)	something (某些事情)	cheesesteaks (干酪牛肉)
paid in full (电影《全款交收》)	yosemite (约塞米蒂)	careing (照顾)	wall street journal 《华尔街日报》

(續表)

白人男性	黑人男性	亞裔男性	拉丁裔男性
mos def talib (茅斯·达夫与塔利·魅力)	feynman (费因曼)	writing (写作)	alternatively (或者)
mangas (日本漫画)	coppola (科波拉)	and my daughter (我的女儿)	mistborn (小说《王者之路》)
abt (关于)	wind up bird (发条鸟)	having (拥有)	weber (韦伯)
utada (宇多田光)	kar (山谷)	brown hair (棕色头发)	gravitate toward (倾向于)

在上面這幾個列表中，拉丁裔的內容最讓我驚訝。長期以來，人口統計工作者經常把拉丁裔美國人和白人併為一類。美國人口普查局在過去多年間也曾嘗試過將他們區分開，但也只是停留在調查問卷的複選框上，在現實中仍然無法將二者區分開。拉丁裔美國人最常用詞表和最不常用詞表為我們揭示了其文化的兩個極端。前者讓我們看到了拉丁文化最深處的音樂與語言，而後者似乎讓我們看到了美國中西部地區大量食用玉米的白人的文化特徵，而這個地區的白人文化幾乎沒有受到過拉丁裔的影響。此外請注意一點，在亞裔最不常用的詞中，要麼存在拼寫錯誤，要麼是普通的職業，要麼是其他一些成就感較低的事情，比如單身父親。

女性列表的內容也同樣豐富和具有啟發性（見表10—5），我仍然建議讀者們仔細推敲一下每一個詞。通過我這種數學算法來分析，亞裔女性最不常講的話包括「我是單身母親」，黑人女性最不常講的話包括

「晒黑」。這些分析結果與我們熟悉的事實完全一致，因此，我不得不說，這種算法是非常準確的，也是我引以為豪的一點。

表10—5 不同族群女性使用最少的詞

白人女性	黑人女性	亚裔女性	拉丁裔女性
fiipino (菲律宾人)	belle and sebastian (贝尔和塞巴斯蒂安乐队)	bbw (肥美女性)	midwestern (中西部)
neo soul (新灵魂乐)	tanning (晒黑)	god my children (天哪, 我的孩子)	cincinnati (辛辛那提)
musiq (音乐)	bruins (棕熊)	single mother of two (两个孩子的单身母亲)	classically (古典)
slow jams (慢拍舞步)	tahoe (塔霍湖)	grandson (孙子)	kenya (肯尼亚)
rich dad poor dad (《富爸爸, 穷爸爸》)	simon and garfunkel (西蒙和加芬克尔)	god my daughter (天哪, 我的女儿)	neal (尼尔)
corinne bailey rae (肯妮·贝儿·雷伊)	magnetic fields (磁场)	mother of three (三个孩子的母亲)	shanghai (上海)
bailey rae (贝儿·雷伊)	sf giants (旧金山巨人队)	human services (社会服务)	financial services (金融服务)
salsa bachata (莎莎巴恰塔舞)	flipping molly (弗根莫利乐队)	degree in criminal justice (刑事司法学位)	classically trained (经过一流训练)
aaliyah (阿莉娅)	head and the heart (乐团名)	single mom of two (两个孩子的单身母亲)	southern belle (南方的美人)
jpop (日本流行)	dodgers (道奇队)	notice my eyes and (看我的眼睛)	cutting for stone (《双生石》)
smh (《悉尼先驱早报》)	wavy (卷曲的)	wanna know just ask (想知道, 随口一问)	in new england (在新英格兰)
salsa merengue (莎莎梅伦格舞)	naked and famous (原装万人迷乐团)	mexican and chinese (墨西哥人与中国人)	antarctica (南极洲)
nujabes (濂叶淳)	social distortion (畸世乐团)	they are my world (他们占据了我整个世界)	kavalier (卡瓦利尔)
48 laws of power (《权力的 48 条法则》)	mountain biking (山地自行车运动)	being the best mom (是最好的母亲)	full disclosure (全面披露)
musiqsoulchild (音乐顽童)	portugal. the man (“葡萄牙·人”摇滚乐团)	raising my children (养孩子)	gravitate toward (倾向于)

(續表)

白人女性	黑人女性	亚裔女性	拉丁裔女性
neyo (黑人歌手尼欧)	camera obscura (暗箱)	a better life for (改善……的生活)	brussels (布鲁塞尔)
2ne1 (韩国 2NE1 女子组合)	rancid (腐臭摇滚乐队)	associates degree in (大专文凭)	toronto (多伦多)
esperanza (埃斯佩兰萨)	yo la tengo (优拉糖果乐团)	curly hair and (卷发)	march madness (疯狂三月)
mangas (日本漫画)	paddle boarding (立桨冲浪运动)	madea (马蒂亚)	cambridge (坎布里奇)
zane (上帝恩典)	armin (阿曼)	im a single mom (我是一名单身母亲)	adventures of kavalier (卡瓦利尔的冒险)
n.e.r.d (书呆子)	santa cruz (圣克鲁兹)	mexican and italian food (墨西哥与意 大利食品)	creole (克里奥尔语)
coldest winter ever (《最冷的冬天》)	ecuador (厄瓜多尔)	i'm a country girl (我是一名乡村女 孩儿)	meetup (偶遇)
mines (矿山)	ccr (跨文化恋爱)	ellen hopkins (艾伦·霍普金斯)	parentheses (括号)
ratchet (棘齿)	the dog park (狗狗公园)	people notice my eyes (人们会注意到我 的眼睛)	arbor (阿伯)
aventura (多米尼加乐队名称)	bbqing (烧烤)	my name is ashley (我叫阿什利)	curl up with a (用……卷起来)
malcolm x (马尔科姆·艾克斯)	origami (折纸手工艺品)	brittany (布列塔尼)	for my next meal (为了我下顿饭)
asians (亚裔)	handshake (握手)	at a daycare (在日间托儿所)	singer songwriters (歌手兼作曲家)
carne (肉)	gabriela (加布里埃拉)	my family my cell (我的家庭, 我的 细胞)	ann arbor (安阿伯市)
hw (如何)	line is it anyway (《台词落谁家》)	want a man that (想找一名……的 男性)	raleigh (罗利市)
earphones (耳机)	sunblock (防晒霜)	me and my son (我和我儿子)	interpreter of maladies (《医生的翻译员》)

到現在為止，我討論了很多與種族有關的現象。正如之前所說的那樣，我之所以這麼做，是因為種族問題一向高度敏感，人們幾乎無法對其進行深入的分析。我收集的這些數據對於研究不同種族的禁忌問題是非常理想的。與種族相比，性別是人類最重要的分類方式，自從人類社會形成伊始，就存在男女之別。由於這種深刻的歷史淵源，性別角色是最具有普遍性的，也最根深蒂固。因為種族界限非常難以清除，所以人們容易忘記這樣一個事實，即種族是時間和地域的產物。愛爾蘭人和東歐人直到20世紀頭10年才被視為「白人」。^[5]在墨西哥，在數百年的時間裡，原住民瑪雅人和具有西班牙血統的混血兒一直是不同的種族，二者也是政治對手，然而，在大部分美國人看來，這兩類人卻被統稱為「拉丁裔美國人」。^[6]由此可見，種族界定會隨著時間和地域而發生改變，但性別的界定卻是固定不變和無法抹去的，在任何一種人類文化和任何一個歷史時期都是如此。

矛盾的是，如果要探索男女之間的差異，OkCupid網站不是最好的數據庫，至少無法用我在這裡提出的方法獲得理想的分析結果。因為這是一個約會網站，你輸入性別的目的只是為了讓其他用戶知道你在尋找哪個性別的朋友，與性別相關的其他信息並不多，用戶所寫的自我陳述也是由其自身性別決定的，存在大量重複的信息，分析效果不甚理想。就分析性別差異而言，最理想的數據應該不受性別的影響，用戶在發佈信息時，不會刻意去顧慮自己的性別。因此，我選擇Twitter作為我的數據來源。在表10—6中，我用的分析方法與分析OkCupid數據的方法是一樣的，只是數據來源不同。

表10—6 男性和女性使用最多的詞

男性	女性
good bro (好兄弟)	my nails done (我修指甲了)
ps4 (索尼发布的下一代电视游戏机 PlayStation4 的简称)	my sissey (我的姐妹)
james harden (詹姆斯·哈登)	mani pedi (同时修手指甲和脚指甲)
mark sanchez (马克·桑切斯)	my makeup (我的化妆品)
my beard (我的胡子)	my purse (我的钱包)
cp3 (克里斯·保罗)	girls night (女孩之夜)
in 2k (电子篮球游戏名称)	my hair for (我的头发)
bynum (拜纳姆)	prom dress (舞会礼服)
the squad (篮球队)	girls day (女孩日组合)
bro we (我们兄弟)	retail therapy (购物疗法)
manziel (约翰尼·曼泽尔)	thanks girl (谢谢姑娘)
in nba (在美职篮)	my future husband (我未来的丈夫)
year deal (签……年的协议)	to dye (染头发)
iverson (艾弗森)	dress shopping (买衣服)
yeah bro (好的, 兄弟)	too girl (太清纯了)
kyrie (姬莉叶)	happy girl (幸福的女孩儿)
hoopin (进篮)	bobby pins (发夹)
free agent (自由职业球员)	wanelo (一个兼有社交和购物功能的网站)
tim duncan (蒂姆·邓肯)	my boyfriend and (我的男朋友)
scorer (得分者)	my belly button (我的肚脐)
offseason (休赛)	my roomie (我的室友)
hof (拜仁霍夫足球俱乐部)	girlies (姑娘)
xbox one (微软开发的家用电子游戏机)	dying my (我快撑不住了)
david stern (大卫·斯特恩)	cute texts (俏皮的文字)
yds (年轻的民主党人)	girl crush (女孩之间的欣赏)
fantasy team (梦之队)	my boyfriends (我的男朋友)
gameplay (游戏玩法)	eyebrows done (修眉毛)

(續表)

男性	女性
gasol (保羅·加索爾)	curl my (卷頭髮或眉毛)
lbj (勒布朗·詹姆斯)	my hubby (我的老公)
bro u (兄弟)	us girls (我們女孩)

通過這個列表，你可以窺見男女兩性的重大差異，讀一遍，你可能會感覺很鬱悶。但在你沮喪之前，要記住我們這個分析方法的目標是找出不同類別的人的獨特之處，也就是說找出一類人有而另一類人沒有的元素，並將其呈現出來。我們並不是用畫筆簡單地勾勒這些人群的輪廓，而是用數學算法對其進行深刻的剖析，展現出其根深蒂固的一些特徵。

表10—6列出的這些詞語具有極為鮮明的性別特徵，屬於極端的情況，但對於男性和女性而言，甚至對於之前提到的不同種族而言，「比薩」等一些基本的詞都是使用頻率非常高的。事實上，儘管當前頗受歡迎的宇宙學將男性和女性置於不同的行星，認為「男人來自火星，女人來自金星」，但心理學家日益形成這樣一個共識，即男女兩性本質上是非常相似的。羅徹斯特大學的研究人員最近也宣稱：「男人來自地球，女人也來自地球。」他們還得出了這樣的結論：「針對13301人的同理心、性取向、科學傾向、性格外向程度等122項指標進行的統計分析表明，男性和女性大體上並不屬於不同的類別。」^[7]

然而，雖然研究人員認為男女兩性本質上大抵相當，但男女差異還是非常顯著的。我的數學算法梳理了多類人群的差異，很難想象哪兩類人群具有男女兩性這樣顯著的差異。我在這裡也不知道應該支持哪一方，因為一方面，女性過度注重外貌，而男性沉迷於球賽、電子遊戲與吃喝之類的事情，那麼世界註定會變得更糟；另一方面，如果男女兩性的生活習慣完全一樣，那麼生活就會失去很多樂趣。同樣的道理也適用於上面提到的不同種族。文化差異雖然有時候顯得荒謬可笑，但世界也

因此變得更加豐富多彩。

火星與金星的說法雖然是一種比喻，但它卻讓我意識到這樣一個事實，即在漫長的科學發展歷程中，宇宙自古以來都是一個參照點。亞里士多德通過觀察浩瀚空曠的大氣來驗證自己提出的以太理論。^[8]牛頓通過觀測火星運動來印證平方反比定律，提出了物體或粒子的作用強度隨距離的平方而線性衰減，即作用力與距離平方成反比關係。1919年的一次日全食證實了愛因斯坦廣義相對論的正確性，才使得愛因斯坦聲名鵲起，因此，可以說沒有太陽和月亮的「幫助」，愛因斯坦的理論就不會獲得真正重要的地位。我們在本書中探討的問題並沒有這些科學人物從事的工作那麼宏大，但不得不說「男人來自火星，女人來自金星」有一定道理，道出了男女兩性在喜歡什麼、談論什麼以及如何打發時間等方面的差異。不過，我希望人們客觀地看待這些差異。有些人提出要努力消除差異，其實這種看法是不成熟的。我們可以這麼想：如果只有地球，而沒有其他星球，那麼宇宙就會非常無趣。

^[1] 正如我在本章展示的那樣，我們可以從一個族群成員的自我介紹文本中總結出這個族群的獨特之處，這個方法是我自己創造出來的。我這個方法的靈感源於OkCupid博客發表的「The Real Stuff White People Like」一文，不過該文采取的總結方法卻是我與馬克斯·施龍（Max Shron）和阿迪蒂亞·穆克吉（Aditya Mukerjee）共同創造的另外一種方法不同。我和他們二人合作的時候，負責在最後階段將無用的數據剔除出去。如果沒有之前的這段經歷，我現在也不會創造出本書採用的方法。本書採用的總結方法由計算機完成，不需要人工介入。我將單詞和短語按照其百分位數放到圖表上，然後根據它們與你選定的那個角的級和距離加以排序。只有在極少數情況下，由於列表中同時出現了一些高度重複的詞，比如「my blue eyes and」「blue eyes and」和「my blue eyes」，才需要人工介入處理，保留最具代表性的單詞或短語，剔除重複的內容。這樣做不會對詞彙列表造成實質上的改變。我這個方法研究的短語都是4個英文單詞以內的，並且出現在30多篇自我介紹文本之中。考慮到空間佈局，我會將三個過長的詞縮短一些，以免自動換行影響對齊。在男性的那個列表裡面，我列的是「follow me」，而不是「follow me on instagram」，在女性那個列表裡面，我用「malcolm x」替代「biography of malcolm x」。在下一章裡關於不同性取向的列表中，我用「feminie women」替代「attracted to feminine women」。

^[2] 我當時已經熟悉了冪律分佈，但還是參考了維基百科「Zipf's law」條目的內容，以及C.約瑟夫·索雷爾於2012年11月5日在《應用語言學百科詞典》（The Encyclopedia of Applied Linguistics）上發表的「Zipf's Law and Vocabulary」一文。

^[3] 另外一個比較有名的巧合例子是歐拉恆等式，即 $e^{i\pi}+1=0$ ，它將數學裡最重要的幾個常數聯繫到了一起。其中包括：兩個超越數——自然對數的底 e 和圓周率 π ；兩個單位——虛數單位 i 和自然數的單位1，以及數學裡常見的0。歐拉真是個天才。

^[4] 這個案例改編自威靈頓維多利亞大學的C.約瑟夫·索雷爾（C. Joseph Sorell）撰寫的《齊普夫定律和詞彙》（Zipf's Law and Vocabulary）一書。這個定律是一個非常好的描述框架，也是經過時間檢驗的。但你可能發現最終的結果與理論所說的存在一定差距，這是正常的，因為這就像你明明知道每次拋擲一枚硬幣出現正面的概率是50%，但你即便拋擲了1000次，也不大可能出現正面的次數剛好是一半的情況。

[5] From Nell Irvin Painter's *The History of White People* (New York: W.W.Norton, 2010).

[6] 小時候，我在墨西哥生活過幾年，之後也一直關注墨西哥的政治動態。See Ronald Loewe, *Maya or Mestizo?: Nationalism, Modernity, and Its Discontents* (Toronto: University of Toronto Press, 2010).

[7] See Bobbi J. Carothers and Harry T. Reis, 「Men and Women Are from Earth: Examining the Latent Structure of Gen-der,」 *Journal of Personality and Social Psychology* 104, no.2 (2013): 385—407. 「Men Are from Mars Earth, Women Are from Venus Earth」 is the title of the article's précis: sciencedaily.com/releases/2013/02/130204094518.htm.

[8] 在寫這本書時，我早就知道愛因斯坦和牛頓在觀察天空的過程中提出了自己的新理論，但我不知道亞里士多德提出以太理論的事情。因此，我在維基百科上找了很久才找到一個我喜歡的例子，請參考維基百科「Aether (classical element)」條目。

第十一章 你墜入愛河了嗎？

幾年前，麻省理工學院的幾名學生利用Facebook的數據開發了一個行之有效的「同性戀雷達」。^[1]這是一款簡單的分析軟件，通過分析一個人在社交網站上的好友信息，從而對其性取向做出合理的預測。這款分析程序可以分析網絡用戶的各種私密數據，判斷其在各個方面的特徵。根據這款程序，如果一個人社交圈裡的同性戀好友和異性戀好友達到了某個特定的比例，那麼就準確無誤地揭示了其性取向。自始至終，不需要了解任何與分析對象直接相關的信息。正如《波士頓環球報》所評論的那樣：「事實上，人們在虛擬網絡中的社交圈可能暴露出了自己真實的一面。」這些學生用自己知道的一些人來檢驗，用這款軟件分析他們的社交圈，判斷他們是不是同性戀，結果表明這款軟件預測的準確率竟然高達78%。這真的非常了不起，畢竟這款軟件僅僅分析一個人的朋友圈信息，而沒有分析與這個人自身直接相關的情況，屬於「盲猜」。我之前猜測這款軟件分析的準確率應該非常低，比如10%、2%或8%，但它居然高達78%，的確相當成功了。

事實上，美國同性戀者的數量究竟有多少，誰也無法真正確切地知道，之前的評估往往相差很多，這也是這些學生開發這款軟件的一個原因。1948年的《金賽性學報告》試圖通過科學方法得到一個準確的數字。^[2]這本書引用和分析了很多案例，提出10%的男性和6%的女性是同性戀。在之後的研究中，有的使用調查數據，有的卻使用在實驗室裡憑空捏造出來的數據，而且有很多是基於政治動機。^[3]有的研究認為同性戀者的比例低至1%，有的認為高達15%。^[4]現在，有了大數據，我們可以通過不同的方法進行更好的、更加準確的評估。提高評估的準確性是非常重要的，因為正如一項研究坦率指出的那樣，「這項工作可以為公共政策的制定過程提供有益的借鑑。」^[5]自1952年以來的歷次總統選舉中，哪怕只有5%的選民改變主意，那麼除了4次總統選舉之外，其他選舉的結果都將是另一番情景。因此，同性戀群體在全體國民中所佔比例究竟是1%、5%還是10%，即便對於政治領域的影響也是非常大的。同性戀者的數量是多還是少，並不會在道德或權利上對其造成任何傷害，即便整個美國只有一個同性戀者，那麼他（或她）仍然享有與其他人同

等的權利，但這個群體的數量無疑會對決策過程產生重大影響，這是一個簡單且不爭的事實。

在歷史上，同性戀群體一直沒有得到過正視。他們長期被貼上汙衊性的標籤，不得不把自己掩飾起來。如果這個群體得到了人們的理解和支持，那麼他們就會表達出自己真實的心聲。他們會大膽地說出：我在這兒。無論如何，同性戀群體都算得上一個非同尋常的少數群體。如果他們認為有必要，就會把自己偽裝成異性戀，至少在表面上會給人這樣一種印象。這樣一來，他們就必須在自我表達與自我保護之間做出痛苦的抉擇，而其他人則無須做出這種選擇。^[6]這種朦朧的、隱蔽的狀態會導致人們對待同性戀的傳統態度和偏見不會受到任何挑戰，進而會永遠維持下去。這會導致同性戀者付出很大代價，同時也會給我們的社會造成諸多損失。因為人們缺乏寬容的態度，同性戀群體就被迫隱藏了自己真實的一面。這樣一來，會催生一種憤世嫉俗的邏輯，即在一個社會裡，如果一個龐大的群體得不到認可，那麼他們就會更加容易走向社會的邊緣，進而產生憤世嫉俗的情緒。然而，從另一方面來講，如果社會認識到他們的存在，就會逐漸學會認可和接納他們。即便保守估算，同性戀者的比例也不會低於2%，這與天生擁有金髮的人群所佔的比例大致相當。^[7]事實上，同性戀似乎比金髮人群更為常見。只是人們不大願意接受他們的存在，因此往往迫使他們偽裝起來，離開人們的視線，其中包括眾多名人。因此，當你下一次拿起一本名人雜誌時，請想一想上面的人物是不是同性戀者。

以大數據為支撐的谷歌趨勢為我們揭示了人們能想而不能說的事情，再次展現了其強大的力量。根據谷歌研究人員斯蒂芬斯—達維多維茨的說法，在美國，關於色情內容的搜索中，有5%是搜索所謂的「關於男同的描寫」，其中有些人直接輸入「男同色情」（gay porn）等赤裸裸的關鍵詞進行搜索，有些人則輸入「火箭筒」（美國一個頗受歡迎的同性戀門戶網站名稱）等相關詞語進行搜索。^[8]更重要的是，無論對於哪個州而言，5%的比例都適用，這意味著男性對同性的慾望不受周圍政治和宗教環境的影響。像這種各州之間在這個比例上的一致性至少為我們提供了以下幾點重要的暗示。第一，它挑戰了「同性戀傾向絕對沒有遺傳性」的觀點。要知道，在環境迥異的密西西比州和馬薩諸塞州，搜索男同性戀內容的比例竟然大致相當，這一證據有力地表明外部力量對同性吸引力幾乎沒有什麼影響。

第二，各州之間在數據上的一致性不僅揭示了與男同性戀者有關的

信息，還揭示出了各州人們在對待同性戀者時，普遍不具有寬容的態度，而這種褊狹的態度在不久的將來就會消除。2013年年初，數學天才納特·西爾弗（Nate Silver）通過其著名的投票模擬算法評估了一下美國各州對待同性戀婚姻的態度。他利用這一算法曾經準確地預測了2008年美國總統大選的結果。他彙總各種數據之後，分析出了各州的公眾輿論態勢，並進行一系列綜合性和前瞻性的分析，然後做出合理的猜測。西爾弗估計，到2020年，同性戀婚姻將在美國44個州獲得合法地位。

如果我們將西爾弗對於同性戀問題的分析結果與美國各州人們在蓋洛普調查中主動透露的性取向做個對比分析，結果是非常有趣的。圖11—1概括性地反映了各州主動表示為同性戀者的人數與西爾弗對於各州對同性戀婚姻的認可度。我的依據是各州法律對待同性戀婚姻的態度，同時著重指出了幾個情況比較特殊的州。

西爾弗的觀點體現在橫軸上，你可以看到，在對待同性戀婚姻的問題上，密西西比州的忍耐度最低，羅得島州的忍耐度最高。縱軸上是蓋洛普的數據，同性戀人群在各州人口總數中的比重最低的是北達科他州的1.7%，比重最高的是夏威夷州的5.1%。從那條傾斜的趨勢線可以看出，一個州對同性戀行為的接納程度越高，那麼該州主動透露自己是同性戀者的人數越多。值得注意的是，如果你順著這條趨勢線向右上側看，當同性戀婚姻獲得100%支持的時候（這是統計學上的一種理想情況，未來，人們對同性戀婚姻可能完全不會存在任何芥蒂，完全能夠接受），你會發現，那個時候，由於已經不再承受任何社會壓力，所以，大約5%的人口會主動承認自己是同性戀者。這個數字與谷歌搜索引擎在不考慮同性戀者的社會壓力的情況下為我們揭示的數字是一致的。

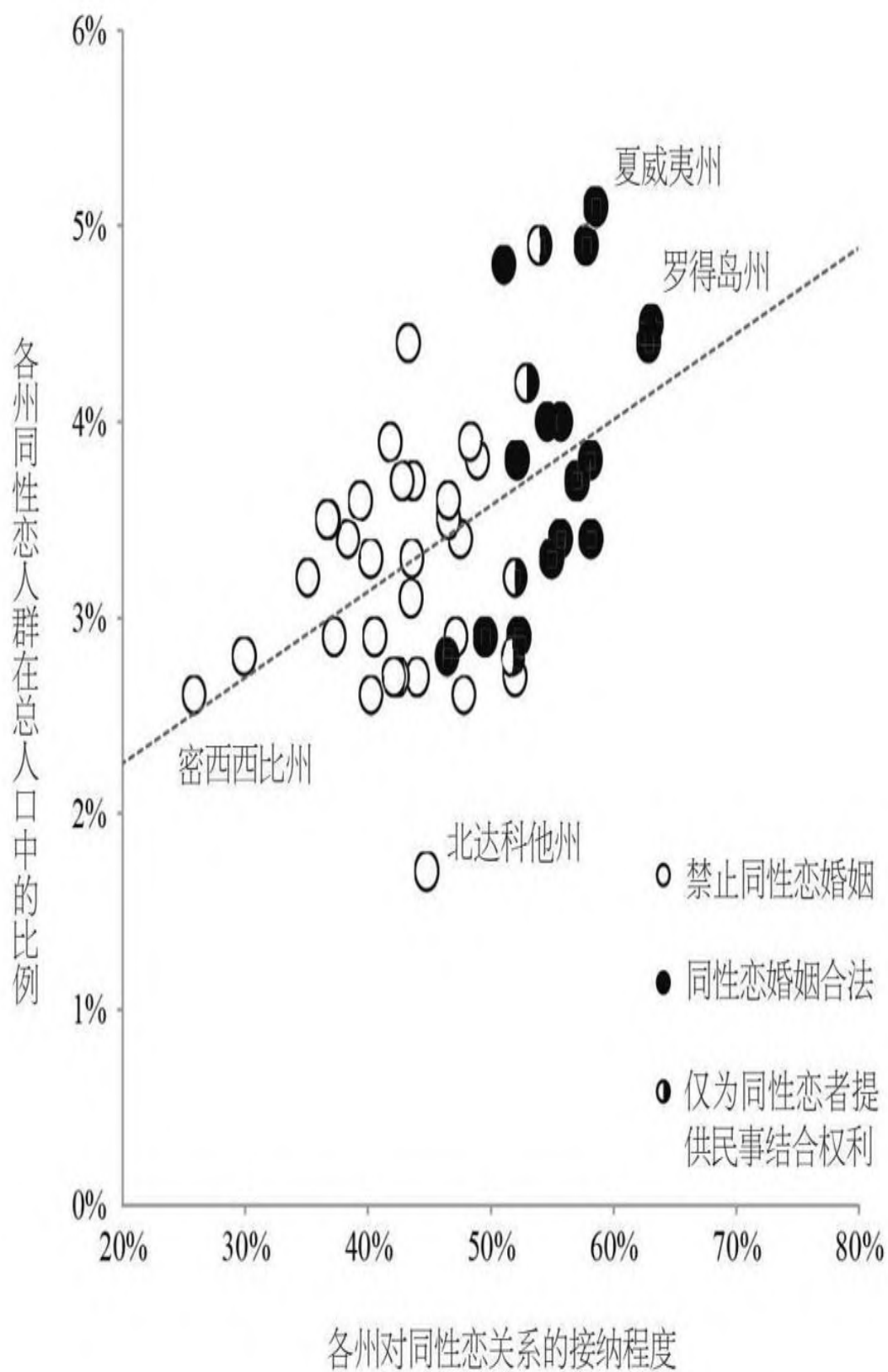


圖11—1 美國各州同性戀人口比例及各州對同性戀關係的接納程度

此外，這條趨勢線表明，同性戀人群並不僅僅居住於對同性戀接納程度較高的地方。我們在前面提到過這方面的證據，即在美國各州關於色情內容的搜索中，關於男同的搜索大約佔5%，這個數值在各州之間都是相近的。此外，我們也可以利用Facebook驗證一下同性戀者的遷移情況，比較一下他們的出生地和當前的居住地，你就會發現，同性戀者很少會因為出生地對同性戀的接納程度低而遷移，同性戀者不會大批大批地搬到更加寬容的地方。^[9]一方面，這體現了家庭關係、家庭撫養教育和生活習慣的力量。另一方面，這也意味著雖然有少部分同性戀者為了追求完整的生活而遷移到了舊金山或紐約之類的城市，但仍然有數十倍的同性戀者選擇壓抑自己，繼續居住在那些對同性戀者容忍度較低的地方。

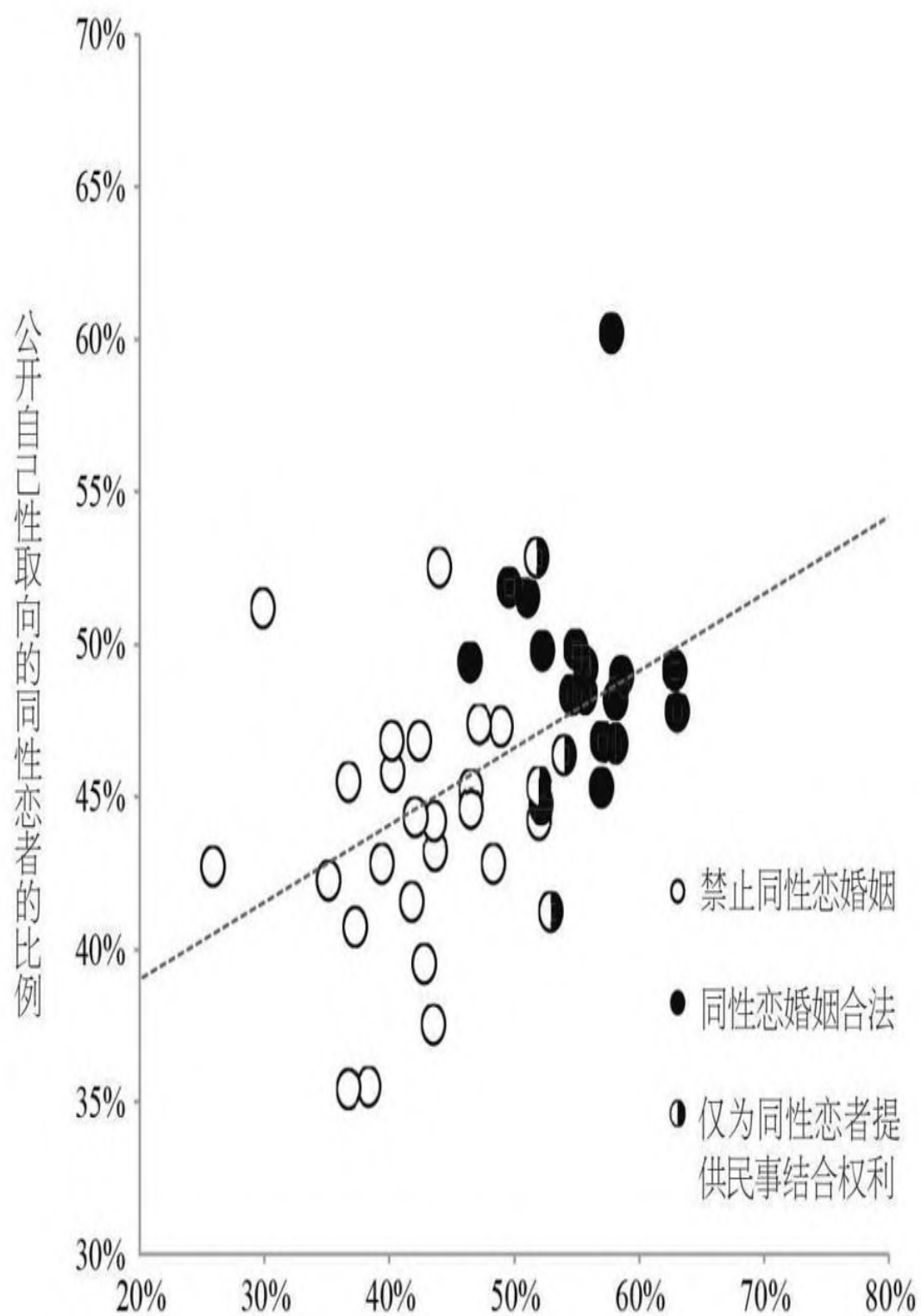
納特·西爾弗、谷歌和Facebook堪稱現代大數據技術中三支重要力量。我們通過這三支力量的分析結果知道，美國每個州的同性戀者在該州總人口中所佔比重大約是5%。如果你認可這個數據，那麼將其與傳統的蓋洛普調查結果進行對比之後，你就會用不同的眼光去看待那些主動透露同性戀傾向的人所佔的比例。比如，如果蓋洛普調查結果告訴我們北達科他州1.7%的人口是同性戀者，那麼你就可能會想到該州大概還有3.3%的同性戀者不願意承認自己的性取向。在紐約，大概4%的人口公開表示自己是同性戀者，那麼或許另外還有1%左右的同性戀者保持沉默。其他州也是這種情況。考慮到同性戀者在美國各州人口中所佔比重的穩定性（5%）以及主動透露同性戀傾向的人口在各州人口中所佔比重的差異性，我們就會得到一個新的結論，即在美國這個國家，有一大批同性戀者生活在祕密狀態下。這就像美國作家、詩人梭羅曾經說過的那樣，「大多數人在安靜的絕望中生活，當他們進入墳墓時，他們的歌還沒有唱出來。」^[10]這些人都是靈魂的難民，我們可以從這些數據中看到他們的存在。

大數據還為我們揭示了由此造成的一些附帶危害。下面，我們再次引用谷歌公司數據科學家斯蒂芬斯—達維多維茨的話：

在美國，如果你在谷歌搜索框中輸入「Is my husband...」（我的丈夫是.....嗎？）進行搜索，那麼系統自動匹配的單詞中，最常見的一個就是gay（同性戀）。在這類搜索的自動匹配中，gay一詞的出現頻率比位居第二位的cheating（欺騙）一詞多出10%，比an alcoholic（一個酒鬼）多出8倍，比depressed（抑鬱）一詞多出10倍。

在同性戀者自我壓抑最嚴重的州，這些疑問式的搜索是最常見的。比如，在南卡羅來納州和路易斯安那州，這類搜索是最多的，而且在這類搜索最頻繁的25個州里面，21個州對同性戀婚姻的接納程度低於全美平均水平。不知道那些堅決逼迫同性戀者轉入「地下」或者堅決「治療」同性戀行為的人如何看待這些數據，如何看待同性戀者被迫與異性結合起來卻缺乏性愛的婚姻，如何看待同性戀者與異性配偶的冷淡關係對孩子造成的負面影響。傳統上，經濟學領域在計算「痛苦指數」時，考慮的是失業率和通脹率。^[11]我建議社會學領域在計算「痛苦指數」時，考慮一下社會中有多少人生活在無法充分地表達自我、無法做真實的自己。這種情況只會給人們帶來痛苦，不會帶來任何好處。^[12]

不幸的是，利用谷歌搜索評估女同性戀者的數量時，效果似乎不太理想，因為很多性取向正常的男性也會去搜索與女同性戀有關的色情內容，從而容易導致我們利用谷歌搜索的數據進行分析時出現偏差。然而，我們在OkCupid網站的數據中可以看到西爾弗關於美國各州接納程度的影子，其中有些差異是很有趣的。我估計，在2013年，美國超過1/4的同性戀者會使用OkCupid網站約會。^[13]一般來說，選擇網絡交友的同性戀者往往比較開放，不介意公開自己的性取向，畢竟，他們只需要在網上寫一份個人資料就足夠了。然而，考慮到很多同性戀者不願意讓所有用戶都看到自己的性取向，所以，網站為這些同性戀者提供了一個隱藏性取向的選擇，他們只要點擊「隱藏」這個按鈕就可以了，只有同為同性戀者的用戶能夠看到他們的性取向。59%的男同性戀者和53%的女同性戀者利用了這個選項。從女同性戀者的數據中，我們也能看到美國各州對同性戀的接納程度與同性戀者的公開程度之間存在明顯的聯繫。接下來，我繪製了圖11—2來表示這種聯繫。



美国各州对同性恋关系的接纳程度

圖11—2 美國各州對同性戀的接納程度與同性戀者公開程度的關係

除了性取向問題之外，OkCupid網站的同性戀用戶與其他用戶看起來非常相似。在網站給出的匹配問題中，同性戀用戶在是否吸毒、是否存在種族偏見以及對性生活的追求與異性戀者處於同一水平，而且同性戀者也渴望同樣的人際關係。事實上，關於對性的態度，如果說有哪個群體是例外的話，那就是異性戀的女性。她們是相對保守的。根據OkCupid網站的數據，6.1%的男異性戀者、6.9%的男同性戀者以及7.0%的女同性戀者都會公然接受隨意的性愛，而只有0.8%的女異性戀者會接受這類性愛。這個數據或許最能說明異性戀的女性在性愛問題上的保守性。^[14]

上述4類人群的性伴侶數量基本上是相等的。男同性戀者與女異性戀者的性伴侶數量的中位數是4個，女同性戀者和男異性戀者的性伴侶數量的中位數是5個，但從總體上看，這4類人群的數據是基本相同的。^[15]如果說在性行為方面有什麼重大區別的話，那就是部分男同性戀者存在濫交的問題，極端的情況是有些男同性戀者主動承認自己擁有25個甚至更多的性伴侶。這類濫交的男同性戀者的數量與濫交的男異性戀者的數量之比為2:1。有趣的是，如同在財富和語言方面一樣，我們在性方面也發現了一個類似的現象，即2%的男同性戀者發生了28%的男男性行為。

為了看一看同性戀者與異性戀者的身份認同感究竟是如何形成的，我們可以按照上一章裡所用的方法來分析這些用戶在網站上所寫的自我陳述，對常用的詞進行排名。正如之前所說的那樣，用戶的自我陳述可以讓我們瞭解各個群體與眾不同的地方，比如女同性戀者有什麼特殊之處以及男同性戀者與男異性戀者有什麼差異等等。這個方法就是根據用戶自己的常用語來分析用戶的一些信息。我們在前面提到的行為數據主要揭示了不同群體表達愛的方式，因此不存在重大差別，但表11—1的一些數據卻揭示了不同群體「愛誰」的問題，因此存在很大差別。我的數學算法幫助我們遴選出了每類人群最常用的詞。

表11—1 各類人群最常用的詞

男同性恋	女同性恋	男异性恋	女异性恋
first wives (前几任妻子)	old lesbian (年龄大的女同性恋)	knows what she wants (知道她想要什么)	honest man (诚实的人)
velvet rage (《天鹅绒之怒》)	i'm a lesbian (我是女同性恋者)	i have no kids (我没有孩子)	man to share (可以 分享……的男人)
tales of the city (《都市奇情》)	i am a lesbian (我是女同性恋者)	treat a woman (请一个女人吃饭)	to meet a man (与一个男人见面)
you're a nice guy (你是个好男人)	femme side (女性同性恋 中充当女方的角色)	care of herself (照顾她自己)	a man who knows (一个知道……的 男人)
anything on bravo (精彩电视节目)	attracted to women who (对……的女人感兴趣)	never been married (没结过婚)	care of himself (照顾他自己)
music madonna (麦当娜的歌曲)	lesbian friends (女同性恋者的朋友)	daughter family (女儿的家庭)	meet a man who (见一个……的男人)

(續表)

男同性恋	女同性恋	男异性恋	女异性恋
music britney (布兰妮的歌曲)	are femme (在女同性恋 关系中充当女方角色)	for a good woman (为了一个好女人)	find a man who (找一个……的男人)
ltr oriented (愿意长期同居的)	butch femme (男性化的女人)	treat a lady (请一位女士吃饭)	meet a man (见一个男人)
romy and michelle's (《阿珠与阿花》)	lesbian movies (关于女同性恋的电影)	good women (好女人)	man who knows how (知道如何……的 男人)
new guys (新手)	single lesbian (单身的女同性恋者)	my kids my family (我的孩子和家人)	a nice guy who (一个……的好男人)
barefoot contessa (《赤足天使》)	u haul U-haul (只约会一 次就同居的女同性恋关系)	hello ladies (《你好女士》)	honest guy (诚实的男子)
kathy griffi (凯西·格里芬)	butch but (打扮男性化, 但……)	type of girl (某类型的女孩)	a man who has (有……的男人)
single gay (单身男同性恋者)	are feminine (有女人味儿)	woman that can (能……的女人)	are a nice guy (是个好男人)
the comeback (《归来记》)	femme who (……的女同性恋者)	real woman (真正的女人)	christian man (信基督教的男人)
hiv positive (艾滋病 病毒测试呈阳性)	elena undone (《艾琳娜的新生》)	my son family (我儿子的家庭)	like a man who (喜欢……的男人)
density of souls (《灵魂集中营》)	the butch (充当男性角色 的女同性恋者)	woman to share (能够分享……的 女人)	a guy who has (有……的男人)
modern family glee (现代家庭的喜悦)	not butch (不充当男性 角色的女同性恋者)	intelligent woman (聪明的女子)	man that knows (知道……的男人)
ab fab (《绝对了不起》)	movies imagine (影片猜想)	god my kids (保佑我的孩子)	love jesus (爱耶稣)
most gay (很有男同 性恋者的气质)	music brandi (布兰迪的歌曲)	girl that i can (我可 以……的女孩儿)	a man who will (一个将能……的 男人)
christopher rice (克里斯托弗·赖斯)	walls could (会……的墙壁) ^①	meet a woman who (见一个……的 女人)	man that has (有……的男人)

a walls could出自電影《如果這些牆會說話》（If the Walls Could Talk），講了女同性戀者的生活片段，產生了強烈的社會反響。——譯者注

（續表）

男同性恋	女同性恋	男异性恋	女异性恋
muriel's wedding (《穆丽尔的婚礼》)	lesbian romance (女同性恋的浪漫)	have no children (没有孩子)	true gentleman (真正的绅士)
other gay (其他男同性恋者)	femme women (在同性恋中扮演女性角色的女同性恋者)	son family (儿子的家庭)	you are a gentleman (你是一位绅士)
flipping out (激动)	debs (女同性恋电影 《少女特工队》)	with the right woman (与合适的女人)	guy to share (分享……的男人)
find mr (找到正确的男人)	feminine women (有女人味的女人)	treat her (请她吃饭)	nice guy who (……的好男人)
guy to date (可以约会的男人)	you're femme (你很有女人味)	right lady (合适的女士)	like a guy who (喜欢……的男人)
sordid lives (同性恋 电影《肮脏的人生》)	soft butch (打扮男性化, 但内心温柔的女同性恋)	great woman (好女人)	a guy that can (能……的男人)
stereotypical gay (典型的男同性恋者)	my future wife (我未来的妻子)	a woman who can (一个能……的 女人)	christian woman (信仰基督教的女人)
flight attendant (空乘人员)	hunter valentine (情人猎 手女子摇滚乐队)	nice woman (好女人)	for a good guy (为了一个好男人)
are you there vodka (《我用青春买醉》)	lesbian looking (看起来 具有女同性恋者的气质)	i like a woman (我喜欢一个女人)	you're a gentleman (你是一位绅士)

如同之前一樣，我建議你深入解讀這些用戶常用的詞，併為你指出幾個總體趨勢。男異性戀者和女異性戀者列表中的詞都與自己當前或未來的伴侶有關。女異性戀者的所有常用詞講的都是她要尋找的那位男性，男異性戀者的絕大部分常用詞講的都是他要尋找的那位女性，唯一的例外就是聊自己有沒有孩子。這些列表讀起來似乎有點類似於電影《人猿泰山》裡的主人公自我介紹時的經典對白「我是泰山，你是簡」（Me Tarzan, You Jane）一樣。如果讓尼古拉斯·斯帕克斯（Nicholas Sparks）去改編一下，或許能寫出一部關於浪漫愛情的小說。

女同性戀的常用詞列表更內斂，自我描述比較多，但仍然非常類似於異性戀者的常用詞列表。女同性戀者尋找的關係和女異性戀者尋找的關係是一樣的，比如，女同性戀者也會讚揚其伴侶很有女人味，或者將其伴侶稱為未來的妻子，只不過具體用詞不一樣。

男同性戀者的常用詞列表與其他三個存在很大區別，因為裡面充滿了與流行文化有關的詞語，而且與直系親屬有關的詞比較少。「精彩電視節目」就很典型，非常恰當地概括出了這類人的偏好。由此可見，在上述4類群體裡面，男同性戀者最不關注性愛和性別認同，我們也可以說，他們的身份認同感除了來自性方面之外，還來自其他方面。

我這種以數學算法為基礎的統計分析是為了凸顯這4類群體之間的差異，但其他一些數據卻表明這些群體之間的界限並非完全固定不變。OkCupid網站向那些在註冊之際選擇「異性戀」的用戶提出過下面這個問題，用戶對這個問題的回答是非常有趣的。

問：你和同性的人有過性接觸嗎？

表11—2 對「你和同性的人有過性接觸嗎？」的回答

	女性		男性	
有过，我很愉快	22 308	26%	12 070	7%
有过，我不喜欢	6 153	7%	10 100	6%
没有过，但期待	14 896	17%	7 632	5%
没有过，也不期待	42 286	49%	137 455	82%
	85 643		167 257	

從表11—2所知，51%的女性和18%的男性曾經與同性發生過性接觸或者期待這類接觸。這些數字比真正的同性戀者評估的數據高得多。因此，我們發現性取向並非固定不變的，而是可變的，網站為用戶提供的性取向選擇可能無法將所有可能性包括進去。從上面的數據中我們可以看到，無論人們對同性性行為的看法如何，這種現象都是比較常見的。

上述數據來自注冊時選擇「異性戀」的用戶，但在關於性取向的下拉菜單中，OkCupid網站還提供了另外一個選項，即「雙性戀」。大約8%的女性用戶和5%的男性用戶選擇了這個性取向。很多人認為雙性戀並非一個真正意義上的性取向，他們認為那些表示自己是雙性戀的人其實是同性戀，只不過他們自己還沒有從心理上真正認同和接受這個事實，只能以雙性戀的名義來掩飾自己。通過OkCupid網站以及其他途徑，我知道很多雙性戀者對這種觀點非常沮喪和不滿。匹茲堡大學公共衛生學院最近的一項研究說得非常好，只是有點直白：「承認自己是同性戀的調查對象對於雙性戀者的反應也不積極……這表明即便在同性戀的小群體中，雙性戀者仍然面臨著嚴重的歧視。」

埃塞克斯大學的傑拉爾夫·裡格爾（Gerulf Rieger）與美國西北大學和康奈爾大學的心理學家在2005年的一篇論文得出結論，從生殖器官對

外部刺激的反應來看，幾乎所有宣稱自己是雙性戀的男性都是同性戀，一些是異性戀，而能夠同時被兩性喚起性慾的人寥寥無幾。^[16]他將男性的雙性戀描述為解讀性喚起的一種方式，而不是能同時被兩性喚起性慾。這一說法激怒了雙性戀群體，這是不難理解的。裡格爾後來重新研究了這個問題，得出了這樣的結論，即男性雙性戀可能是一種「好奇心」，想看看別人的裸體、觀察別人做愛、看色情電影或參與性愛派對，他認為這是男性的一種興趣。因此，從這一點來看，就不難解釋為什麼那些自稱能夠同時被兩性喚起性慾的人實際上只能被一個性別的人喚起性慾。他們在思想上對兩個性別感興趣，而他們的身體卻會區別對待兩性。

通過分析OkCupid網站上的數據，我們似乎找到了一些證據來支持裡格爾的結論，因為大多數雙性戀用戶只是給一個性別的人發送信息。在圖11—3中，我展示了雙性戀用戶發送信息的情況。

我在這裡所說的只給某個性別發信息，指的是一名用戶將至少95%的信息發送給了這個性別的用戶。因此，這個標準是很高的，在統計過程中不存在任何馬虎的地方。從圖11—3的數據可以看出，只有一小部分所謂雙性戀用戶給兩個性別的人發信息。這種情況似乎印證了裡格爾的說法，即有些人雖然宣稱自己是雙性戀，但這並不能反映出其真實行為。有趣的是，男性並非一成不變地給某個性別發送信息，有時候會將大量信息發送給某個性別，而有時候卻會發給另一個性別。比如，那些宣稱自己是雙性戀的年輕用戶中，超過一半的人只給其他男性發送信息，而年齡越大的雙性戀用戶，給異性發送信息的情況越少，35歲左右的男雙性戀用戶只給女性發信息。很多男雙性戀者隨著年齡見長，逐漸坦然接受了同性戀的事實，而不再宣稱自己是雙性戀。從這種變化中，我們似乎能看出，的確有一部分男同性戀者用雙性戀來掩飾自己的性取向。但要真正深入地研究同性戀者性取向的變化，需要沿著時間跨度去分析，目前，我們還沒有這樣的縱向數據。

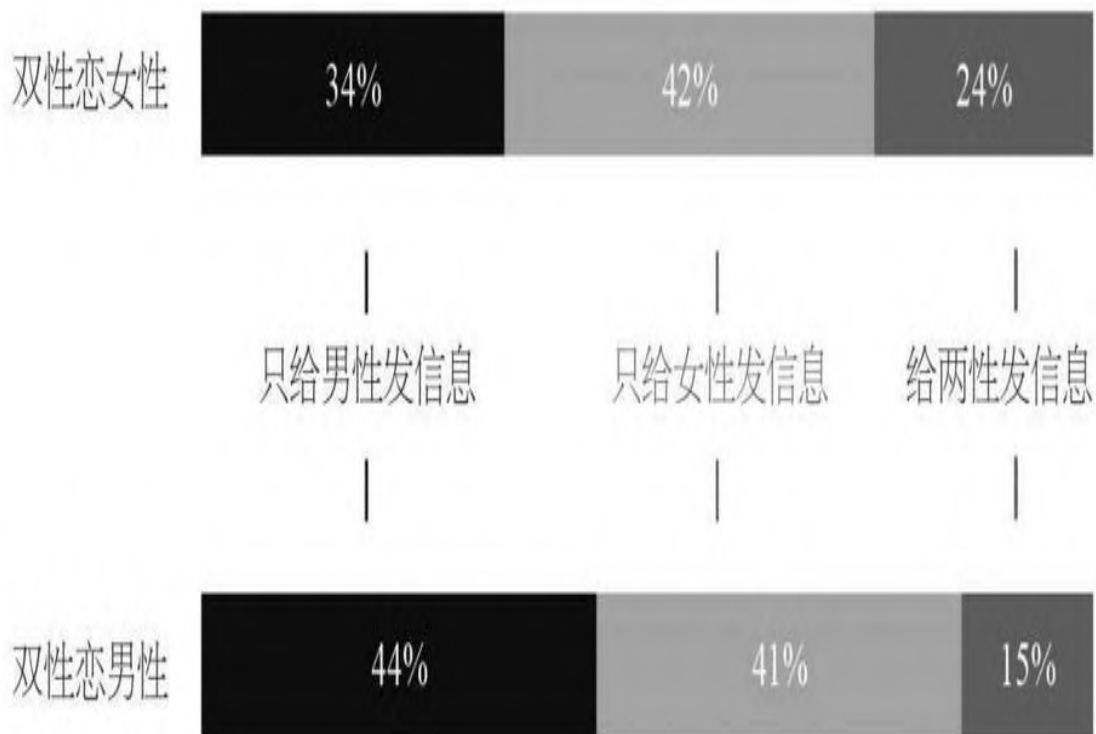


圖11—3 雙性戀者發送信息的情況

由此可見，我們口頭上說自己是什麼樣的人與實際行為其實是兩碼事，後者不應該自動地否定前者。畢竟，人們應該享有按照自己的方式來描述自己的權利，如果一味地要求人們去迎合一名研究人員（或一個網站）的定義，去迎合他人所貼的標籤，其實是沒有意義的。如果實際情況與標籤存在出入，那麼出錯的一方肯定是標籤，因為人們有權去愛他們覺得正確的事物，有時候用來描述現實狀況的語言可能跟不上現實的變化。比如，在2014年情人節那一天，Facebook給出了50多個性取向的選擇，而不是隻能選擇男性或女性，甚至給用戶提供了「變性人」和「兩性人」的選項。波士頓雙性戀資源中心總裁埃琳·魯斯特倫（Ellyn Ruthstrom）認為：「不過，這些研究將性和性關係僅僅侷限於性刺激。研究人員想把雙性吸引格式化——你必須與定義的一模一樣，對男性和女性都感興趣，那才是雙性戀——這太荒謬了。我認為人們應該有權通過多元化方式向外界證明自己的雙性戀傾向。」^[17]作為OkCupid網站的管理者，我在雙性戀用戶的自我陳述文本中的確能夠發現這種多元化的表達方式。30個最常用的單詞包括「雙性戀」（bisexual）、「泛性戀」（pansexual）、「異性裝扮癖」（cross-dressing）、「有同性行

為的異性戀」(heteroflexible)。還有一些雙性戀者在描述自己的狀態時會說喜歡與家人在一起，喜歡自己的工作。這些話可能表明他們在現實生活中時常得不到理解與接納，時常被視為「異類」，因此感到非常孤單，也很不滿。

女雙性戀者的情況略微不同於男性的情況。對於女性而言，雙性戀和同性戀的情況似乎更加常見，而且並不像男同性戀那樣遮遮掩掩，至少美國女歌手麥莉·賽勒斯(Miley Cyrus)的情況能說明這一點。可能是因為那些從事宣傳推廣工作的人深知性問題的炒作有利於擴大市場，而且明星們需要掙脫傳統的束縛，變得更加前衛。因此，同性戀問題受到了持續關注和炒作，今天的流行文化中充斥著這類元素。雖然我並不是非常確切地瞭解麥莉·賽勒斯的情況，但她的同性戀問題得到熱炒之後，她所穿的衣服似乎都實現了空前銷量。就像吉恩·西蒙斯(Gene Simmons)的同性戀問題得到熱炒之後，面部彩繪開始變得非常流行一樣。相似地，很多針對年輕男性的網絡騙子也會拿性問題作為掩護，他們在社交網站上註冊虛假賬號的時候，常常會選擇雙性戀作為自己的性取向。在Facebook上，58%的虛假自我陳述都說自己是「女雙性戀者」，相比之下，在真實的賬戶中，宣稱自己是女雙性戀者的人只有6%。^[18]在OkCupid網站上，問題並沒有這麼明顯，但如果用戶選擇雙性戀以及其他幾個關鍵指標，肯定會引起網站管理者的格外關注。

絕大部分自我描述文本都是真實合法的。但如果研究一下這些文本，你就會發現，很多女性宣稱自己是雙性戀者，有些女性甚至公開地邀請其他女性同自己的男朋友或丈夫組成三口之家，很多男異性戀者也時不時地幻想自己是同性戀。如果你仔細分析一下，就能發現這一點（見表11—3）。

如果我把上面這些女雙性戀者常用詞譜上曲子，讓美國當紅嘻哈歌手皮普保羅(Pitbull)去演唱一番，肯定朗朗上口，勇奪第一。說了這麼多，雖然通過炒作性取向而實現商業計劃的做法非常愚蠢，但最起碼主流社會不再刻意壓制這些取向了，而是認識到了其中可以利用的價值。事實上，對於性取向而言，我們看到情況正在發生變化，而且變化速度很快。納特·西爾弗通過我們前面提到的數學算法，發現了美國社會在過去10年間的一個重大轉變，即美國社會對於同性戀婚姻接納程度的轉變在2004年明顯加快了。他指出，之前人們只是樂觀地假定同性婚姻支持者或許在不久的將來會佔據全國大多數，而他根據自己的數學算法預測，這種假定很快就會變成現實。

因此，用數學算法來分析，我們不難發現情況正在逐步好轉。同性戀問題長期以來都存在，早在19世紀晚期就有很多同性戀者，人們當時將主動透露性取向視為一種政治行為。^[19]「出櫃」這個詞語也是多年之後出現的。現在，經過同性戀者經年累月的努力呼籲，他們公開地生活和示愛的目標已經基本上實現了。不僅很多名人紛紛出櫃，還有數以百萬計的普通人也大方地公開自己的性取向，雖然我可能永遠不知道這些普通人的姓名，但他們無疑在一定程度上提高了社會對同性戀者的接納程度。在不久的將來，民意調查者會放下蔑視同性戀者的姿態，科學家會用另一種視角看待同性戀問題，雄心勃勃的學者也可以將數學算法用於計算其他事情，而不僅是計算隱晦的同性戀問題。終將有一天，世界會變得非常開放，我們再也不需要去絞盡腦汁猜測了。

表11—3 女雙性戀者常用詞

bi female (女双性恋)
bisexual female (女双性恋)
me and my husband (我和丈夫)
me and my man (我和我的男人)
my boyfriend is (我的男朋友是)
hubby and (丈夫和)
we are a couple (我们是情侣)
i am bisexual and (我是双性恋)
me and my boyfriend (我和男朋友)
fun couple (快乐的夫妻)
couple we (我们俩)
married couple (夫妻)
we are not looking (我们不寻求)
fun with me and (我觉得很有趣)
do have a boyfriend (有一个男朋友)
my bf and (我男朋友和)
female to join (女性加入)
girl to join (女孩加入)
another couple (另一对情侣)
bi woman (双性恋的女人)
my boyfriend (我的男朋友)
i am bi sexual (我是双性恋)
my hubby and (我丈夫和)
join me and my (加入我和我的)
female for (为了……的女性)
my boyfriend and i (我男朋友和我)
we are looking to (我们正想)
a triad (三口之家)
no single (非单身)
send us (给我们发……)

[1] Here, I used 「Project 'Gaydar,'」 by Carolyn Y. Johnson, Boston Globe, September 20, 2009, and the students' original paper, 「Gaydar: Facebook Friendships Expose Sexual Orientation」 by Carter Jernigan and Behram F.T. Mistree, First Monday 14, no.10 (2009), firstmonday.org/article/view/2611/2302.

[2] 關於這一點，請參考維基百科上的「Kinsey Reports」條目，該條目簡要介紹了金賽針對男女兩性的研究內容。「10%的男性是同性戀」這個數據是比較準確的，但「6%的女性是同性戀」這個數據則存在較大的不確定性。《金賽性學報告》說，純粹同性戀者（Exclusively Homosexual）在女性人口中所佔的比重約為2%~6%。

[3] 關於這一點，請參考維基百科「Demographics of sexual orientation」條目，該條目給出了各個性取向的人口數據。此外，也可以參考維基百科「LGBT demographics of the United States」條目的內容。

[4] 調查數據往往會受到外在因素的影響，比如調查人員在提問時的遣詞造句方式，以及對於性經驗與性別認同感的關係的權衡方式。

[5] Dan Black et al., 「Demographics of the Gay and Lesbian Population in the United States: Evidence from Available Systematic Data Sources,」 *Demography* 37, no.2 (2000): 139—54.

[6] See AssiAzar, 「Op- ed: To You There, in the Closet,」 *The Advocate*, April 16, 2013, advocate.com/commentary/2013/04/16/op-ed-you-there-closet/.

[7] 我的數據來源是賓夕法尼亞州西彭斯堡大學C.喬治·博伊裡（C.George Boeree）教授發表的一篇名為「Race」的帖子，該帖子獲取鏈接為：[web space.ship.edu/cgboer/race.html](http://web.space.ship.edu/cgboer/race.html)。只要做一個簡單的數學計算，就能證明他的觀點。比如，歐洲人、加拿大人、美國人和澳大利亞人加在一起大概有10億人，如果其中1/6的人天生擁有金髮碧眼（在我個人的交際圈裡面，這個1/6的數字是大大誇張了），那麼這就相當於全球人口的2%了。

[8] 為了探討人們搜索男同性戀色情的情況及其意義，我用了長達4頁的篇幅（指英文版中的4頁）。我講的這些內容，借鑑了達維多維茨於2013年12月7日在《紐約時報》上發表的「How Many American Men Are Gay?」這篇文章。我引用的谷歌趨勢的數據、納特·西爾弗和蓋洛普的每週數據都來源於這篇文章。Silver's original piece is 「How Opinion on Same Sex Marriage Is Chang-ing, and What It Means,」 from his New York Times fivethirtyeight blog, fivethirtyeight.blogs.nytimes.com/2013/03/26/how-opinion-on-same-sex-marriage-is-changing-and-what-it-means/. Gallup's numbers are from Gary J. Gates and Frank Newport, 「LGBT Percentage Highest in D.C., Lowest in North Dakota,」 gallup.com/poll/160517/lgbt-percentage-highest-lowest-north-dakota.aspx.

[9] 達維多維茨在其文章裡將研究範圍擴大到了可以公開獲取的Facebook用戶的個人介紹信息。

[10] 我引用的這句話裡面，有一段話來自梭羅的《瓦爾登湖》。此外，我還從奧利弗·溫德爾·霍姆斯（Oliver Wendell Holmes）的詩作《無聲者》（The Voiceless）裡面摘出來了兩行。See The Walden Woods Project: walden.org/Library/Quotations/The_Henry_D._Thoreau_Mis-Quotation_Page.

[11] 請參考維基百科「Misery index (eco-nomics)」條目，最初的計算公式是阿瑟·奧肯（Arthur Okun）提出來的。

[12] 這種情形還有助於某些造成這種情形的人達到其政治、宗教和娛樂目的。

[13] 這基於兩個假設：大約5%的美國人是同性戀；人口普查報告指出的9300萬美國單身人士中，有一半正在試圖尋找約會對象。政府將每一個沒有結婚的人都算作「單身人士」，這在評估真正的單身人口時，特別是考慮到同性戀人群的存在，顯然是有問題的。2013年，

OkCupid上註冊了65萬個同性戀者的檔案，大約佔了積極交友的美國同性戀群體的26.8%。雖然有少數賬號是重複的或很少使用的，但我們的網站在美國同性戀交友市場所佔的份額依然是巨大的。在這個報告中，「同性戀」和「雙性戀」用戶是分開的，這種計算不包括後者。

[14] 有些男同性戀交友軟件是專門為一夜情而設計的，其中最著名的兩個就是Grindr和Scruff。異性戀常用的軟件是Tinder，受歡迎程度不亞於同性戀交友軟件。因此，我認為就OkCupid上的男同性戀人群而言，選擇性偏差不會比異性戀族群更糟糕，但我承認這一點很難準確證實。

[15] 49%的異性戀男性和同性戀女性表示自己的伴侶不超過4個。

[16] 我參考了裡格爾教授及其團隊的兩篇文章：GerulfRieger, Meredith L.Chivers, and J.Michael Bailey, 「Sexual Arousal Patterns of Bisexual Men,」 *Psychological Science* 16, no.8 (2005): 579—84, and its successor, GerulfReiger et al., 「Male Bisexual Arousal: A Matter of Curiosity?,」 *Biological Psychology* 94, no.3(2013): 479—89。

[17] See David Tuller, 「No Surprise for Bisexual Men: Re-port Indicates They Exist,」 *New York Times*, August 22, 2011, and Meredith Melnick, 「Scientific Study Finds That Bisexuality Really Exists,」 *Time*, August 23, 2011, healthland.time.com/2011/08/23/scientific-study-finds-that-bisexuality-really-exists/.

[18] See Chris Taylor, 「Fake Facebook Users Likely to Be Popular Bisexual College Women,」 *Mashable*, February 3, 2012, mashable.com/2012/02/03/fake-facebookusers-bisexual-college-women/.

[19] 關於這一點，請參考維基百科「Timeline of LGBT history」和「Coming out」條目。最初是卡爾·海因裡希·烏爾裡希斯（Karl Heinrich Ulrichs）將同性戀者自我披露性取向（即「出櫃」）視為同性戀群體的權利意識增強的表現。

第十二章 瞭解自己所處的位置

我讀初中時，午餐休息時間很長，因為我們當時年齡都大了，不想做遊戲，不想玩耍，所以吃完午餐之後，就跑到學校後面的那塊空地上聊天，等上課鈴聲響起時，再回教室裡。在七年級剛開學那幾天裡，我們都是三三兩兩地站在柏油路面上，而最初的位置一旦確定下來，之後三年裡基本上沒有什麼改變。我記得，從距離餐廳最近到最遠，排列順序依次是：

- 超酷的孩子們（大多數來自學校所在城市的富人區）
- 預科生
- REM搖滾樂隊和Cure搖滾樂隊的粉絲們
- 滑冰愛好者
- 重金屬樂迷
- 我和我的朋友們
- 一個棕色的大垃圾桶
- 交換生和有學習障礙的學生

顯然，這種排列方式絕不僅僅是偶然性的。那個大垃圾桶自然變成了較為弱勢者的集合地點，在垃圾桶的兩側，學生呈現出明顯的分化。我那一端的孩子們一般都是討論《忍者神龜》等角色扮演遊戲，而不討論電視上播放的兒童類節目，因為這類節目被視為專為小孩子準備的。每個人似乎都是根據某一種根本性的、主導性的力量來決定自己所站的位置。

數字資料的一個美妙之處就是數量非常大，此外，如同我們初中的那塊空地一樣，同時具有物理和社會維度。一張紙有兩個維度，時空具有四個維度。弦理論預測，我們的物理存在大概需要10~26個維度；我們的情感世界肯定具有這麼多甚至更多的維度。如果我們能將自己的內心世界同外部世界聯繫起來，那麼我們對存在狀態的描述就會更加深入。

到目前為止，我們研究人類以及人際互動的方式，也就是研究人際關係、自我描述文本、相互評價等因素的方式，基本上都忽略了物理空間這一因素，即忽視了人們所處的地點，而網站和智能手機具有定位功能，當然能夠為我們收集豐富的位置數據。Twitter上的帖子都標註著相應的經度和緯度，Facebook會主動詢問你的家鄉在哪裡、你在哪裡讀的大學以及你現在住在哪裡。很多應用程序甚至能準確地定位你所處的大樓。接下來，我們將把物理空間與身份認同、情緒、行為和信仰放在一起分析，看看會獲得什麼樣的新認知。我們將看一看地點如何塑造一個人以及人們如何在古老的土地上劃定新的邊界。

很多社會或社區的分界線都是由法令或偶然事件確定的，有的是在這兩種因素的綜合影響下確定的。美國和蘇聯之所以將北緯38度線作為停戰的軍事分界線，把朝鮮一分為二，就是因為當時《國家地理》雜誌刊登的一張朝鮮地圖把這條緯度線突出地標了出來，這條線只花了半個小時就選定了。^[1]在同一個月稍早的時候，德國被劃分為幾個佔領區，依據便是各方軍隊當時的駐紮地點。美國很多州之間的界線是由皇家憲章或美國國會法案確定下來的，而劃界過程的主導者們可能從來都沒有親自踏上過那片土地。這種劃界方式在非洲、印度次大陸、中東以及其他遭遇過帝國殖民統治的地方產生了嚴重的危害，其影響至今仍然存在。只有在很少的情況下，地圖能夠反映出「人民的意志」，但即便如此，也存在不少問題，比如，在以色列與周邊國家的邊界衝突就是一個例證。在世界近代史上，以色列最初是英屬巴勒斯坦託管地。後來一個關鍵問題逐漸凸顯出來，即劃界過程要體現哪些人的意志。

對於網站而言，在收集數據時，需要考慮到與數據有關的政治邊界和自然邊界。在網絡世界裡，信息是流動的、抽象的、沒有邊界的，信息在網絡世界中的流通就類似於貨幣在現實世界中的流通。在信息面前，將現實世界割裂開的武斷的分界線往往是令人討厭的。在OkCupid網站上，河流一直都是對距離匹配算法干擾最大的因素。在紐約市，皇后區與曼哈頓區隔河相望。從物理維度來看，二者相距不到半英里；但從社會維度上看，二者卻存在天壤之別。因此，要努力用計算機能夠識別的方式將這些差異解釋給計算機。當一個人上網的時候，他既是現實世界的一部分，又遠離了現實世界。但反過來想，這種雙重性意味著我們能夠根據一些新的維度去審視固有的物理空間，這些新的維度可能比那些因板塊構造或法令催生的邊界更有意義。

圖12—1是克雷格網站（Craigslist）對美國地圖進行劃分之後的結

果，這個地圖上的每一個小區域都由一個專門的團隊負責。^[2]一位地圖製作者曾經將這幅圖稱為「克雷格合眾國」（the United States of Craigslist）的地圖。但我覺得「合眾」這個詞用在這裡是錯誤的，因為這個地圖明顯將美國地圖分割得更為嚴重，而且在整個地圖內，每個小區域似乎都是一個獨立的小王國，就像歷史上分裂的神聖羅馬帝國一樣。



圖12—1 Craigslist網站對美國地圖進行的劃分
一旦我們在這幅地圖的空間裡填充上具體內容，就變得更加有趣

了。我們不妨深入看一看克雷格開創的這個「帝國」。下面是一個孤獨的人在該網站「尋親覓友」（Missed Connections）欄目上發的一個帖子：

我們一起從34街進入Q住宅區。你穿著一件雙排扣短呢大衣，你的眼睛很有奧黛麗·赫本的風采，我們還相互對視了幾下。如果你看到這個帖子，請給我發電子郵件。

這個帖子描述的故事可能發生在曼哈頓。美國俄勒岡州波特蘭市，人們往往更有可能在公交車上發生眼神交流。在加利福尼亞州，人們則更有可能在健身時進行交流。但對於美國其他地區，最容易發生邂逅、最令人懷念的場所卻是沃爾瑪超市（見圖12—2）。^[3]

現在，我們所看到的內容是傳統製圖師無法做到的，因為衛星發現不了這些內容。上面這種簡單的圖將人類行為與自然地理結合了起來，屬於一種新型的地圖。

在上面的例子中，克雷格網站先劃定邊界，選定自己想要服務的市場。大多數網站都收集有關地理位置的數據，而不是預測。根據這些數據，我們就可以製作另外一種世界地圖，實際上就是根據人類行為來改變固有的邊界線和輪廓線。前幾年，才華橫溢的數據分析家、軟件工程師皮特·沃頓（Pete Warden）根據2.1億Facebook用戶的社交資料，針對美國各城市之間的社交網絡進行了群聚分析（cluster analysis），結果發現美國可以劃分成7個群聚。^[4]他劃分美國版圖的方式非常新穎，其依據不是政治因素，而是人與人之間的友誼。位於同一個群聚內的使用者較容易結交為朋友，但不同群聚之間的使用者較少接觸。這7個群聚分別是：太平洋群聚（Pacifica），社交網絡完全以西雅圖為中心的美國西北太平洋地區；南加州群聚（Socalistan），加州與內華達州的使用者，幾乎都會結交洛杉磯與舊金山的朋友；摩門群聚（Mormonia），猶他州與愛達荷州南部的使用者自成一體，較少與外部來往；牛仔群聚（Nomadic West），在美國西部，使用者傾向與距離非常遙遠的使用者來往；大得州群聚（Greater Texas），以達拉斯為中心，圍繞著墨西哥灣海岸、俄克拉何馬州、阿肯色州的使用者；南方群聚（Dixie），幾乎與美國內戰時南方各州的範圍相符；宅男群聚（Stayathomia），從新英格蘭一直延續至明尼蘇達，這個地區的特色是，使用者多與距離很近的使用者來往。

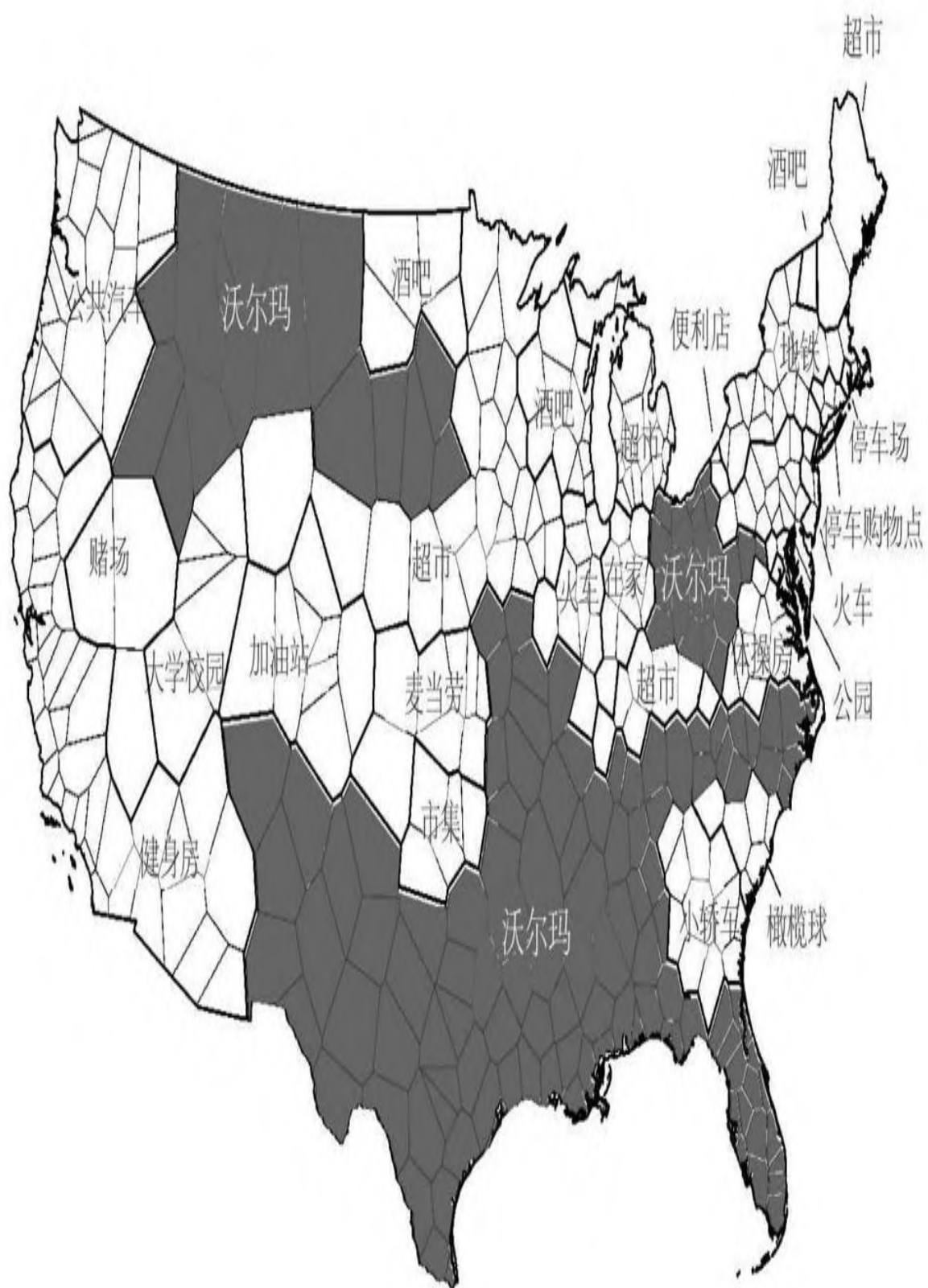


圖12—2 不同地區的人容易邂逅的地點

從那以後，具有GPS（全球定位系統）功能的智能手機為製圖學領域帶來了革命性的變化。肯塔基大學的地理學家馬修·祖克與數據科學家合作創建了他們所稱的「多麗項目」（DOLLY Project, Digital On Line Life and You，意為「你和數字在線生活」）。^[5]該項目儲存了2011年12月以來Twitter上的每一個帖子，每個帖子均帶有相應的地理信息，用戶可以在該儲存庫中自由搜索。這就意味著祖克和他的團隊收集了上百億條帶有具體經度和緯度的帖子，能夠根據這些帖子去審視縱橫交錯的社交網絡。多麗項目是一個用途非常廣的資源，現在人們正在探索它的用途。祖克已經將其用來分析一些具有高度私密性的事情。2012年2月，他在列剋星敦的辦公室遭到了地震破壞。他利用自己的數據庫來分析地震對人們造成的心理影響，圖12—3的地圖顯示了Twitter用戶對這次地震的反應強度。整幅地圖是圍繞這次地震的震中繪製的。圖中的等值線表示驚訝程度相等的點的組合，驚訝程度自內而外逐漸弱化。

祖克發現「情緒震中」恰恰位於「地震震中」（肯塔基州哈澤德市）西北方向不遠的地方。這幅圖雖然看似簡單，其發現卻具有開創性的意義。與其他類型的地圖相比，它屬於一種全新的地圖。比如，克雷格網站的分類地圖在20世紀70年代也能做出來，該網站「尋親覓友」欄目本來就是從報紙上「尋人啟事」欄目演變而來的。因此，在互聯網誕生之前，如果你真的想做，可以蒐集一下本國100個最大城市的日報，記錄一下上面的尋人啟事信息，會發現與我們之前看到的那個尋人帖子沒有多少區別。此外，從理論上來講，即便在幾十年前，只要某個研究小組有足夠多的資源去上門採訪數以百萬計的人，並跟蹤分析一下其社交網絡的重疊部分，就有可能找出美國的宅男群聚，與皮特·沃頓根據Facebook用戶的社交數據劃定的群聚沒有什麼區別。

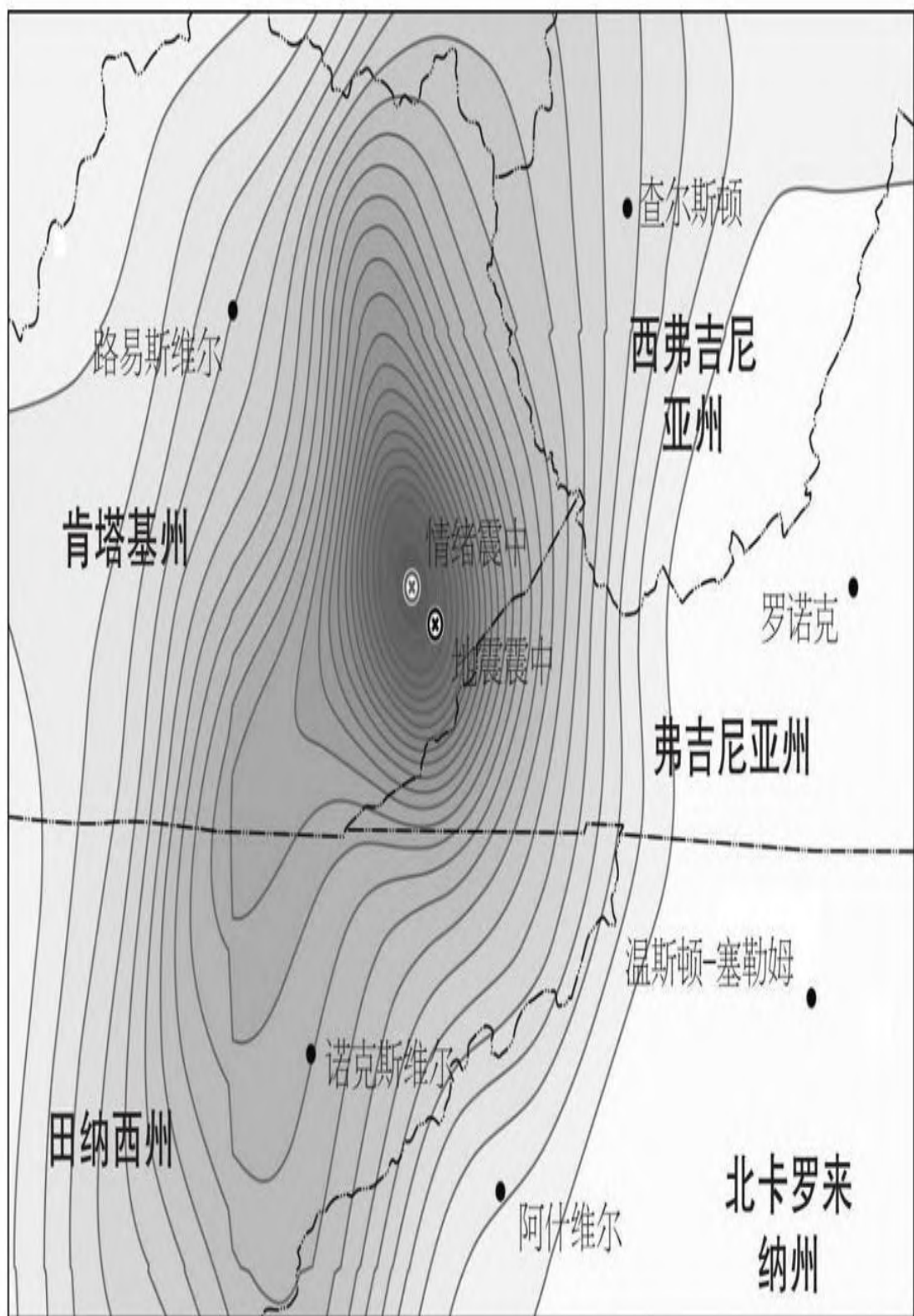


圖12—3 Twitter用戶對一次地震的反應強度

但祖克的地圖卻顯示了人們在地震那一剎那間的反應，整個過程極為短暫，而記錄下的信息卻非常及時和全面。在大數據技術誕生之前，是無法做到這一點的。如果在地震之後去採訪肯塔基州的人們，即便費盡九牛二虎之力，也不可能獲得一個真實的報告，因為人們的情緒和記憶都會出現變化，媒體的報道和人們的談論也會「汙染」數據。智能手機的出現並沒有使地震儀過時，但與傳統的里氏震級圖相比，祖克繪製的情緒變化圖卻以一種更加直觀的方式反映了地震對人們情緒造成的「衝擊」。如果你的工作是向受害者發放援助物資，而你一時又無法瞭解地震造成的破壞程度，那麼上面這種情緒反應圖就能給你提供一個更好的指導。^[6]

雖然每一個帖子都是短暫的，但把海量帖子彙集在一起進行分析，就能得到一些長期性的結論。YouTube網站上有一個視頻展示了「多麗項目」的強大功能。在該視頻中，人們用它來跟蹤荷蘭人在聖馬丁節這一天發的帖子。在德國、奧地利和荷蘭地區，這個節日大體上類似於美國的萬聖節，小孩子們提著燈籠挨家挨戶敲門，站在門口唱歌，索要糖果、點心、水果或其他禮物。你肯定能想到荷蘭北部人口密集的城鎮會慶祝這一節日，在數據中也能看到這一點。但出人意料的是，比利時西部地區也慶祝這一節日。就這樣，Twitter用戶發的帖子把老荷蘭地區同比利時的弗蘭德斯（Flanders）地區聯繫在了一起。之所以出現這種現象，就是因為這兩個地區具有相似的文化基因。因此，雖然那些帶有GPS定位信息的帖子是沒有生命的，但經過深入觀察和分析，我們就能看到哈布斯堡家族的身影。

既然「多麗項目」之類的軟件具有如此強大的力量，而我們卻缺乏能夠拿來分析的縱向數據，這的確是一件令人苦惱的事情。在今天用於研究的數據庫中，雖然數據量非常大，但大體上都是十幾年來的數據，稍早一些的數據則不存在。Twitter的確能夠為我們提供多維數據，讓我們瞭解他人的情緒，幫助我們定位到地球上的每個點，但也僅僅侷限於Twitter出現之後這些年的數據，再早一些的數據就無從查找了。在歐洲歷史上，地理因素、文化因素和語言因素在長達數百年間的互動和融合過程一直變幻不定，比如，阿爾薩斯—洛林地區的主導權就經歷了德國—法國—德國—法國的「易手」過程，每一個政府都試圖將其文化施加給這個地區的人民，導致這個地區看起來就像一個被塗上了層層油漆的房子。我們可以想象一下，如果我們擁有年代足夠久遠的數據，將這個歷史過程記錄下來，那將會是什麼樣的情景？我們也想象一下，如果我

們的數據能夠記錄下15世紀晚期加勒比海地區的歷史事件，我們就能直觀地看到那個時期先後主導過這片土地的士兵、宗教和語言（從阿拉瓦語到阿茲特克語）。「多麗項目」創建的目標就是讓我們看到一個文化的滄桑和破裂，現在需要的只是時間的積累。[\[7\]](#)

除了Twitter上的帖子之外，其他數據也能為我們揭示出其背後隱藏的地域文化，甚至能讓我們從另一個角度獲得深刻的見解，只不過與網帖相比，其他數據在即時性方面會大打折扣。當網站直接給用戶提出問題徵求答覆時，我們不僅有機會以合理的方式調整固有的界線，而且可能會發現其實固有界線與人們平常所想的並不一樣。

OkCupid網站向用戶提出了這樣一個問題：「燒國旗是否違法？」圖12—4彙集了100萬份答案。在這裡，我的製圖軟件沒有畫出任何政治或自然界線，只是分析用戶給出的答案，然後根據這些用戶的緯度和經度將其排列到相應位置。這個國家的確存在不同的原則和不同的觀念，或者說，這個國家可以被劃分為兩個部分：城市部分與農村部分。你甚至可以看到一個地區侵佔了另一個地區：在哈德孫河上游和加利福尼亞州北部「葡萄酒之鄉」的農村社區，是用來自大城市的資金建立起來的，因此這些社區居民給出的答案與大城市居民給出的答案是一致的。



圖12—4 美國各地對「燒國旗是否違法？」的回答

相似地，通過「多麗項目」蒐集的網帖來分析，我們會發現關於同性戀的搜索也是不分州和國家的。這與我們之前根據「谷歌趨勢」得出的「同性戀是普遍現象」的結論是一致的。圖12—5顯示了各地用戶從「海盜灣」網站^[8]下載同性戀色情內容的情況。^[9]這幅地圖沒有參考任何固有邊界，也沒有其他指導因素，完全是根據用戶的IP（網際協議）地址來繪製的。前面那幅關於「燒國旗是否違法？」的地圖為我們呈現了兩種截然對立的答案。相比之下，圖12—5的主題卻是「團結」：從加拿大的埃德蒙頓和卡爾加里往南，一直到墨西哥的蒙特雷和奇瓦瓦，人們表現出了同樣的愛好。我們生存的地方就是這樣。



圖12—5 美國各地從海盜灣網站下載同性戀色情內容的情況

我們有多少個數據來源，就有多少個製圖方法。我們之前一直採用的製圖方法基本上是依據心理維度，即人們對於國旗和色情內容的心理感受，然後依據相關的經緯度信息繪製成圖。但採用另外一種維度也是有可能的，即化抽象為具體。數據可以幫助我們以具體的方式探討一些抽象問題。接下來，我們再次通過OkCupid網站的數據來分析一下人們是否講衛生的問題。圖12—6揭示了廣大用戶的洗澡頻率。

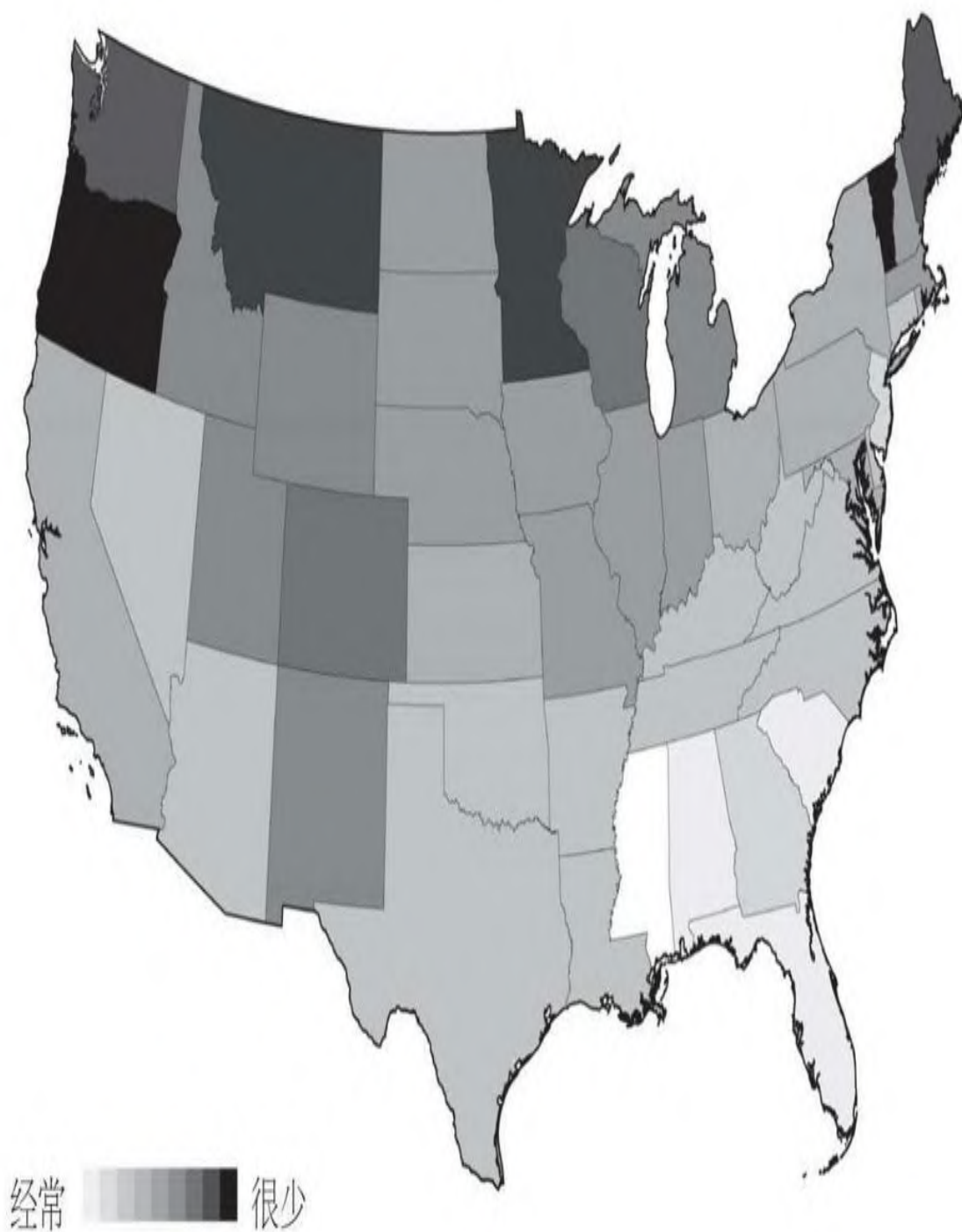


圖12—6 美國各地對「你多久洗一次澡？」的回答

一方面，圖12—6反映出了一個大趨勢，即天氣越熱，人們洗澡的頻率越高。但如果我們深入分析一下細節，就能發現一些比較有趣的現

象。在新澤西州居民給出的答案中，你會看到很多類似於GTL^[10]的字眼。事實上，與周邊其他州的居民相比，新澤西州居民在衛生問題上的確比較挑剔。而在佛蒙特州居民給出的答案中，你卻發現了相反的人生哲學，該州是居民洗澡次數最少的幾個州之一，與相鄰的新澤西州相比可謂是天壤之別。根據谷歌搜索的結果得知，現在佛蒙特州用於體現其形象的動物是摩根馬（Morgan Horse），或許用一個梳著「駭人」長髮絡的白人男子更合適。^[11]

政治、天氣、沃爾瑪以及地震都與現實中的物質世界具有密切的聯繫，但在我們的數據庫中，我們甚至能窺探到人們內心深處的隱私想法與地理因素的關係。我們以性慾為例再次展開分析。從理論上來講，性慾不應該有邊界，各州的情況應該是一樣的；但數據卻為我們揭示出一個令人驚訝的事實（見圖12—7）。

從圖12—7上來看，美國中部偏北以及西部地區在性慾方面更開放、更大膽、更主動。這種情況在OkCupid網站上得到了一次又一次的印證。你可能會認為太平洋沿岸的幾個州在這方面會更多地表現出非傳統的態度，但在這些紅肉消費量較大的州里，好幾個州都不是你所想的那樣。要知道，從「政治傾向」這個指標來看，OkCupid網站在南達科他州和北達科他州的用戶們表現出了符合人們預期的保守性，而且他們在該網站上的自我描述文本與其他州用戶的文本沒有多少區別。此外，從其他指標來看，該州的陰影也不應該像圖12—7顯示的這麼暗。但在這些數據中，我們看到這兩個州的居民在性慾方面卻表現得格外開放。這種出人意料的現象為我們揭示出了互聯網數據的一個神奇力量，即傳統上的社區往往受到地理界線的限制，而藉助互聯網數據，我們卻能發現一些超越地理界線、以其他因素為聯繫紐帶的「社區」。

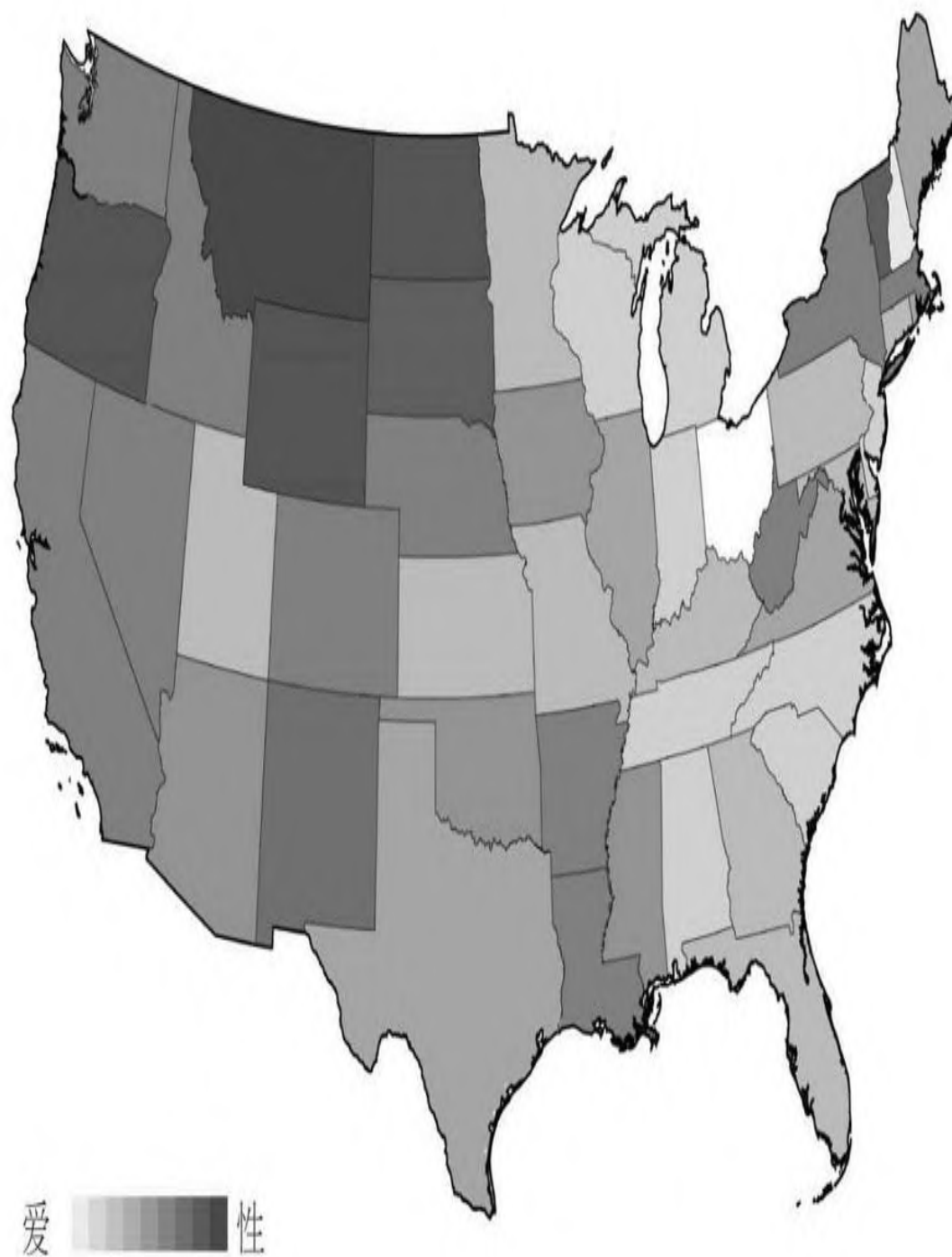


圖12—7 美國各地對「你現在認為性與愛哪個更重要？」的回答

上面這些數據並不意味著這些位於山地時區的城市就是約會者的樂園，原因其實聽起來也很老套：如果你在南達科他州的皮埃爾市尋找約

會對象，但本地選擇非常有限，便不得不求助於約會網站，以至在網上表現得更開放。雖然我們的數據來源主要是這些在現實中無法滿足而求助網絡的人，在數據樣本選擇方面難免存在一定的偏差，但我們最終的數據仍然是很有意義的，因為通過這些數據，我們發現如果人們在現實世界中無法得到滿足，就會到網絡世界中創建電子社區。對於約會網站而言，這類社區就意味著一群在性方面具有相似偏好的人走到了一起。在其他用途較為廣泛的網站上，這就意味著用戶不僅僅是為了三三兩兩地聚在一起調情，而是希望能夠得到更多。

Reddit就實現了互聯網最早的野心，即打破地理空間的制約，將不同地點的人們聚在一起聊天、辯論、分享、傳播新聞和大笑，從而營造密切的人際關係。這是最受歡迎的社交化新聞和娛樂網站之一，^[12]其口號是力爭成為「互聯網主頁」。註冊的用戶可以將在互聯網上搜集或原創的圖片、材料以帖子的形式在網站上發佈，而後其他用戶可以對這個帖子進行投票，結果將被用來進行排名和決定它在首頁或子頁的位置。你在一些大的聚合網站上看到的內容都來自這個網站。我可以舉個例子，而不是在開玩笑。我在寫這本書的時候，《赫芬頓郵報》網站上發佈的一段視頻非常火。該視頻的標題是：「這隻鹿心想沒人注意到它放屁，現在全世界都知道了」。在該視頻中，一隻鹿躲在一棵大樹後面放了一個屁。我可以向你擔保，這段視頻肯定是最先被上傳到了Reddit上，肯定是該網站最先看到這隻鹿放屁的。

奇怪的是，雖然該網站的影響力非常大，但網站本身似乎什麼都沒有做，沒有應用程序，沒有遊戲，更沒有用戶的自我描述文本。網站在紐約的辦公室也是好幾個人共用的，還沒有我的臥室大。該網站上的內容也只是用戶提交的原始鏈接，其他用戶會對其進行投票、評論、修改和轉載，給人的感覺就像世界上最大的朋友群坐在世界上最長的沙發上聊天。該網站的用戶們很少知道彼此的名字，更不用說親自見面，但這種匿名性並沒有妨礙他們建立密切的關係。2011年感恩節的前一天，舊金山灣區一名40歲的女性在該網站上發帖稱自己很孤獨，結果在短短幾小時的時間裡她的帖子就收到了500多條評論（當然，許多人邀請她第二天去赴晚宴）。^[13]帖子的影響迅速擴大，將許多城市的用戶也聯繫到了一起。

該網站具有數以千計的子頁，是依據興趣來分類的。每一個子頁都是用用戶創建和維護的，都有一群忠實的發帖者和評論者。在這個一無所有卻又無限寬闊的網絡空間裡，人們創建了真正意義上的虛擬社區，社

區的主題包括遊戲、科技、音樂、全國橄欖球聯盟（NFL）以及很多只能在該網站上才能發現的主題。比如以下幾類主題就是很常見的：[\[14\]](#)

「求解答」類——「印度教、佛教都說人有輪迴，那麼它們如何解釋人口增長呢？」

「我是.....」類——「我是採訪新澤西州州長克里斯·克里斯蒂（Chris Christie）的記者，有問題儘管問！」

「今天瞭解到」類——「今天瞭解到，俄勒岡州一個名為「無聊」（Boring）的小鎮和英國蘇格蘭地區一個名為「乏味」（Dull）的小村近日結為友城。」

「求助」類——「Reddit上的前煙友們，你們到底如何成功戒菸的啊？」

「誰會贏」類——「至尊超人vs戴著無限手套的超人，哪個會贏？」

在圖12—8中，我列出了200個最流行的主題。這幅圖與我們前面看到的Craigslist網站的那幅圖存在很大的相似性，二者在劃分美國版圖時採用的算法是相似的。Reddit這個圖的劃分依據是用戶的興趣，揭示出了Reddit用戶的整體心理狀態，展現了一些不盡相同卻又相互聯繫的網絡社區。在這幅地圖中，每一個「州」的版圖大小與其主題的受歡迎程度是一致的，我的製圖軟件根據各個子頁上的網友相互評論的情況，將具有共同愛好的網友彙集在了一起。

如果你不熟悉某種組織和展示語言信息的方式，那就像我們之前所做的那樣，先找出幾個熟悉的詞，然後憑感覺猜測一下這些詞為什麼會以這種方式組合在一起。對於我而言，這是很簡單的。我最喜歡的遊戲是「萬智牌」（MagicTCG），其周圍的主題是「男人的權利」

（MensRights）、「誰會贏」（whowouldwin）、「我的小馬」

（mylittlepony）。相似地，很多體育類主題，比如全國橄欖球聯盟、美職業籃球協會（NBA）、一級方程式賽車（Formula 1）等，都位於該圖的最下方。與《口袋妖怪》（Pokemon）有關的主題都位於該圖左側。在該圖右側，「英國的問題」（Britishproblems）挨著「澳大利亞」（australia）和「英式足球」（soccer）。最流行的一些主題位於該圖中間，也就是說，距離其他任何一個主題的位置都不遠。每個主題就是一個子頁，該頁顏色的深淺取決於該頁用戶的密切程度，顏色越深，表示這個子頁越孤立，用戶只在該頁發帖子的可能性就越高，用戶之間

的關係也就越密切。正所謂物以類聚，人以群分。圖12—8也向我們揭示出一旦用戶認為某些事情是有趣的、好笑的或重要的，就會自發地在網絡世界裡彙集到一起，而不是根據他們在現實世界中的生活地點彙集在一起。這是「集體意識」的體現。

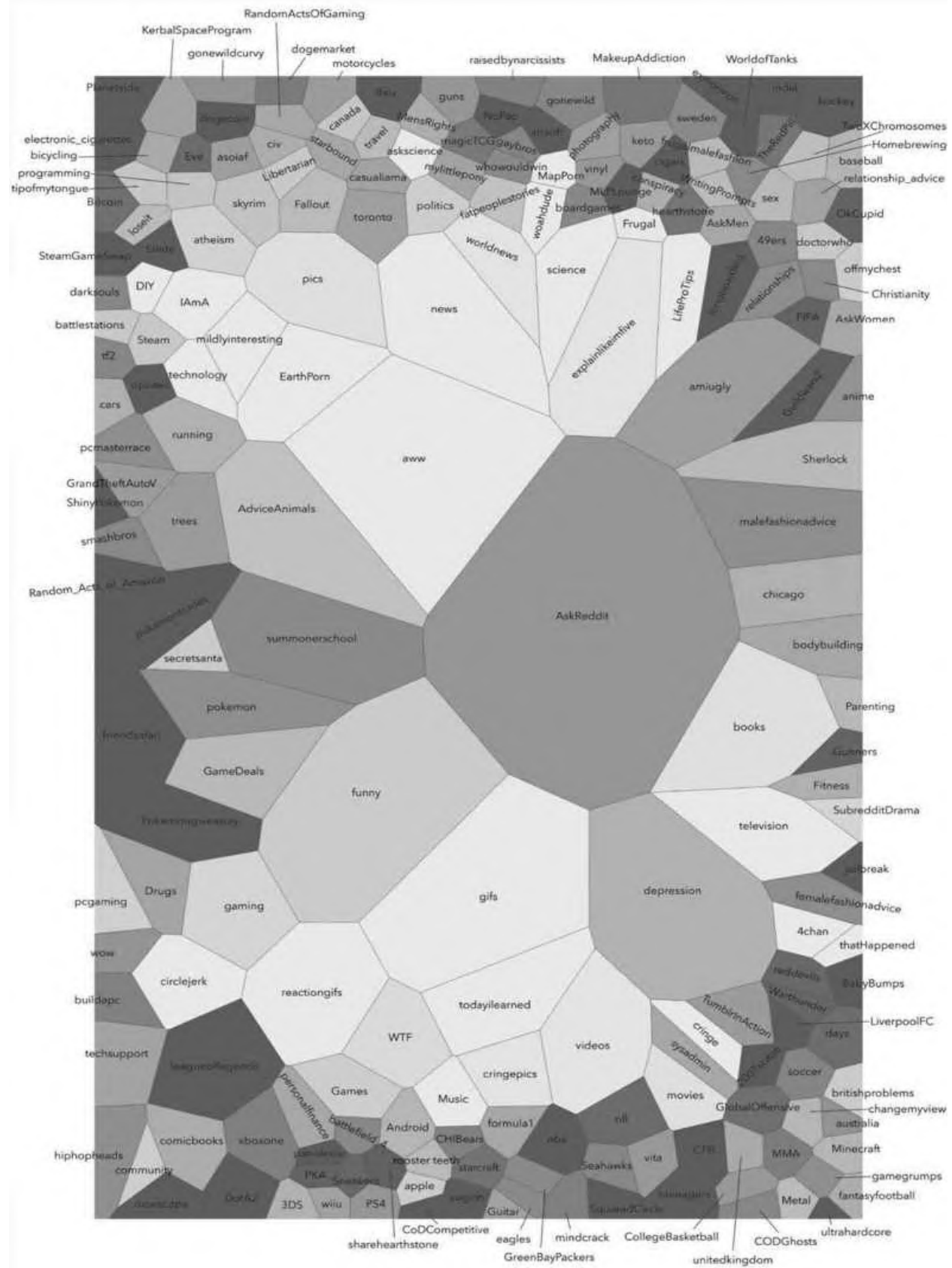


圖12-8 Reddit網站上流行主題分佈圖

本尼迪克特·安德森（Benedict Anderson）是康奈爾大學的教授，他寫的一本書原封不動地在我書架上躺了很長一段時間。我在讀大學期間就應該讀一讀的，不過一直沒有讀。這些年我搬了幾次家，無論走到哪裡，都會把它帶上。這本書的名字是《想象的社區》（*Imagined Communities*）。我翻開讀了讀，因為這個書名似乎與我寫的這本書有些關聯。安德森在該書中討論的主題是民族主義和建立國家，他認為國家是想象出來的，因為即使在最小的國家裡，一個國民永遠不可能認識大部分同胞，不可能遇見或聽說過大部分同胞，但每個人的頭腦中都浮現著同胞的形象。^[15]他這本書寫於1981年，但似乎在談論互聯網。雖然Reddit不是一個國家，但是裡面活躍著很多具有共同愛好的人，這些志同道合的人可能永遠不會見面，但不妨礙他們形成密切的關係。我們在前面講過薩菲亞、娜塔莎、賈斯汀在Twitter上遭遇的網友攻擊，這是集體暴力的表現，具有悠久的歷史。現在，在Reddit上，我們看到了一個圍繞著共同興趣建立起來的新型網絡社區，的確非常有趣。在這些社區成員的互動中，我們看到了國民集體意識中比較美好的元素，包括歸屬感、同情和分享。

現在，我已經在紐約市布魯克林區生活了12年。我從書架上把這本書拿下來的時候，上面佈滿了紐約的灰塵，不過這本書最早跟我去得了得克薩斯州。我大學剛畢業時，和其他幾個人住在一起，其中有一位叫安德魯·布西內斯克（Andrew Bujalski）的老兄決定搬到美國得克薩斯州的首府奧斯汀市，因為他喜歡《茫然與困惑》（*Dazed and Confused*）和《都市浪人》（*Slackers*）這兩部電影，而這兩部電影的導演理查德·林克萊特（Richard Linklater）就住在那裡。因此，可以說那是一次朝聖之旅。我們其餘人也沒有什麼計劃，就索性跟著他一起去逐夢了。現在，布西內斯克已經成了一位知名導演。

當然，像那次說走就走的旅行只是一群22歲的孩子的朝聖之旅，我們沒有更好的事情可做，只能跟著他人去追逐夢想。我們之前只是聽說過奧斯汀這個地方不錯，因此我們就去了。這是一個無足輕重的例子，但正是這些基於口碑和對未來的美好期許的集體遷徙創造了我們所知的世界。20世紀初，數百萬非洲裔美國人離開了飽受歧視的南方，遷徙到了底特律、芝加哥和紐約等北方城市，推動了美國文化的轉變。個人的遷徙或許微不足道，但數百萬人的遷徙就形成了大規模遷徙的洪流。加利福尼亞州經歷的淘金熱是如此；歐洲人最初遷徙到北美大陸，推動新

舊世界的融合也是如此；我想，13000年前大批克勞維斯人穿越冰橋來到北美大陸，成為這片土地上第一批原住民的遷徙運動還是如此。人們遷徙既是為了尋找一個維持生存、獲得安全的環境，也是為了尋找一個自己內心想法被尊重的地方。

最近，Facebook的數據科學研究團隊基於全球視野分析了一下現代社會的大規模遷徙。^[16]他們將其稱為「協調性遷徙」，即一個地區內的大部分居民成批遷徙到其他地區的行為。在美國內部，這類大規模遷徙運動已經不存在了，但在世界其他許多地方，這類遷徙才剛剛開始。谷歌公司的研究人員繪製出了全球遷徙地圖。我在這裡只摘錄了這個地圖中的東南亞部分：圖中的線條表明小城鎮和農村居民大規模地遷徙到中心城市。這幅靜態的遷徙圖（見圖12—9）表明這個地區正在迅速發生變化。無論如何，1850年左右的英國就是這樣，50年之後的美國也是這樣。

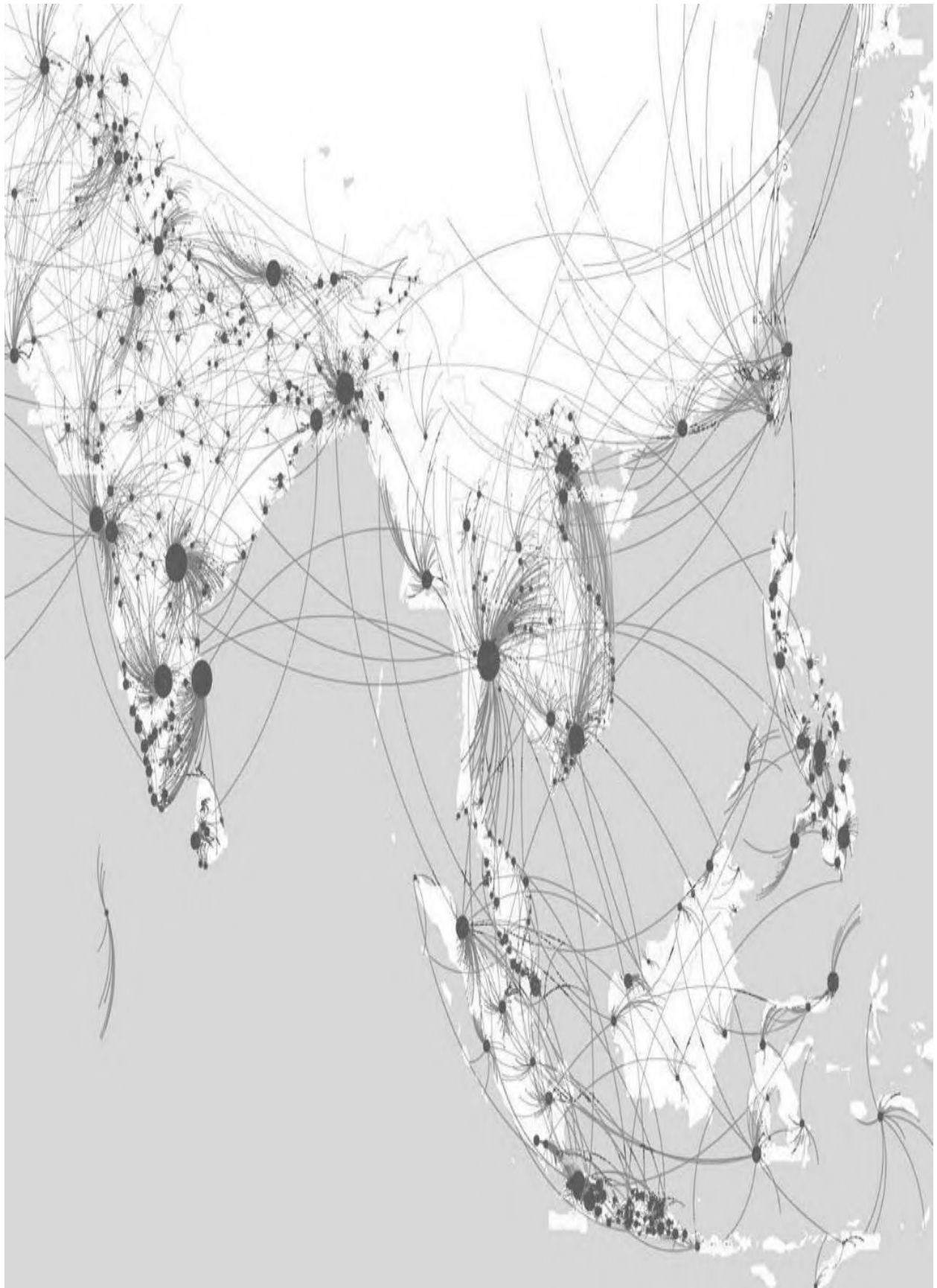


圖12—9 谷歌繪製的全球遷徙地圖

從最廣泛的意義上來講，在這些大規模遷徙背後，最有可能的驅動力是經濟因素，因為在芝加哥或曼谷之類的大城市，工作機會較多，前景較好。雖然這張地圖上的線條和點代表著大趨勢，但對於千千萬萬個遷徙者而言，遷徙的決定無疑具有獨特的意義。是父母決定收拾一下行李就走嗎？有朋友引路嗎？這些人到了新城市之後和哪些人在一起呢？他們在家鄉留下了誰呢？他們把所有東西都帶走了，還是把所有東西都留下了？我也在猜測，每個人遷徙時都會帶一本書，並且最終會去讀嗎？如果帶的話，他們會帶什麼書呢？

[1] 我在寫這本書時對於選定這條分界線的過程已經有了一個大致的瞭解，因為之前我讀過麥克阿瑟將軍的回憶錄《美國愷撒》，但我在本書中提到的這個令人難以置信的逸事則是引自維基百科「Division of Korea」條目。該條目引用的原始資料是唐·奧伯多弗（Don Oberdorfer）所著的The Two Koreas（New York: Basic Books, 2001）。我在谷歌圖書上也搜索出了這本書的內容，證實了這則逸事。該書在谷歌圖書上的獲取鏈接為：books.google.com/books/about/The_Two_Koreas.html?id=yJZKpYXh2SAC。

[2] 如同本章其他美國地圖一樣，這張地圖以及Reddit圖都是由詹姆斯·多德爾製作的。這是對美國進行細分之後製作出來的標準的維諾圖，每個克雷格市場就相當於每個州的首府，在維諾圖中被稱為細胞或種子。雖然這張圖看起來顯得錯綜複雜，但實際上其背後的原理很簡單，也就是在最接近的兩個種子之間，取與二者距離相同的點，然後將這些點連接起來，併成為一條中間線。我見過該圖的多個版本，而我自己所畫的這個版本的靈感來源是IDV Solutions製作的那張圖，該圖由「john.nelson」貼在了UX博客上，獲取鏈接為：uxblog.idvsolutions.com/2011/07/chalkboard-maps-united-states-of.html。

[3] 該圖與之前的維諾圖是同一幅圖，但結合了多夢西·甘布里爾（Dorothy Gambrell）發表在《今日心理學》（Psychology Today）的「尋親覓友」地圖的數據。原始地圖的獲取鏈接為：psychologytoday.com/blog/brainstorm/201302/missed-connections-0。為了同前面那個Craiglist圖保持一致，根據前面那個維諾圖的分割方式對這幅圖的數據進行了處理。

[4] 這個黑客是才華橫溢的數據分析師、軟件工程師皮特·沃頓（Pete Warden）。他發表的文章是「How to Split Up the US」，獲取鏈接為：petewarden.com/2010/02/06/how-to-split-up-the-us/。沃頓後來在另一篇名為「Why You Should Never Trust a Data Scientist」的文章中承認說，他把美國分成7個新區，只是隨便的劃分方法，僅供數據分析界基於娛樂目的使用，我在這裡其實也是發揚了同樣的精神。

[5] 祖克教授及其團隊有一個非常好的地理博客，我對其研究成果的瞭解主要來自這個博客，該博客名為Floating Sheep，鏈接為floatingsheep.org。我對於地震和這張地圖的討論，主要來自泰勒·謝爾頓（Taylor Shelton）發表在Floating Sheep博客上的一篇文章，該文章題為「Mapping the Eastern Kentucky Earthquake」。採用的圖則在下面這張原始圖片的基礎上進行了簡化，以便於印刷：floatingsheep.org/2012/11/mapping-eastern-kentucky-earthquake.html。「多麗項目」的團隊成員包括祖克教授本人以及馬克·格雷厄姆（Mark Graham）、泰勒·謝爾頓（Taylor Shelton）、莫妮卡·斯蒂芬斯（Monica Stephens）和阿特·珀修斯（Ate Poorthuis）。關於荷蘭人在聖馬丁節這一天走家串戶的情景，來自珀修斯的講述，他講述的視頻獲取鏈接為www.youtube.com/watch?v=pD9HWAaQGUA。我對於學生狂歡引發亂局的討論來自於傑里米·W.克拉克頓（Jeremy W.Crampton）等人在《製圖與地理信息科學》（Cartography and Geographic Information Science）期刊上發表的「Beyond the Geotag: Situating‘Big Data’and

Leveraging the Potential of the Geoweb」一文，該文獲取信息為：Cartography and Geographic Information Science 40, no.2 (2013)：130—39。

[6] 兩個月後，祖克又測量了另外一種情緒波動：肯塔基野貓隊贏得了全美大學生籃球聯賽，學生們極度興奮，大肆慶祝。一位名叫@TKoppe22的網友在Twitter上發了一個帖子說：

「哦，這裡有一個半裸的男子，手拿一個煤氣罐#LexingtonPoliceScanner。」之後，#LexingtonPoliceScanner這個標籤開始流行了起來。祖克追蹤了這個標籤，看看之前僅僅侷限於部分地區的標籤如何傳遍世界。在Twitter上，自炫博學與低級趣味並存，好像精神分裂一樣，令我感到震驚。這簡直就是高科技的鬧劇。

[7] 我知道除了需要等待時間的積累之外，還有一個必要條件，即有關人士必須使用Twitter。對於哥倫布時代的美洲而言，這當然是一種奢望。但正如我在前面所說的那樣，現在的Twitter比很多人認為的要普及得多，也更加大眾化和民主化。如果現代社會發生了諸如征服西班牙之類的事件，肯定會被人發到Twitter上。

[8] 海盜灣網站，Pirate Bay，是一個專門儲存、分類及搜尋種子文件的網站，是美國網絡分享與下載音樂、視頻等文件的重要網站之一。——譯者注

[9] IP地址並無法準確地指明一個人（或者更準確地說，是一臺電腦）所在的位置，只能把範圍縮小到10~15英里之內。weather.com之類的網站也使用了類似的技術，你不需要輸入你的位置，網站就能大致猜測出你所在的城市。IP地址只能給我們一個大概的信息。在本段的研究中，我們只知道一臺電腦正在下載哪些色情內容，其他信息則無從得知。此外，我們也不知道是誰在操作這臺電腦，甚至無法確定是不是真的有一個人人在操作這臺電腦，也可能是開了外掛。

[10] GTL，即gym、tan和laundry的首字母縮略詞，意為「健身、晒身和洗衣」，是電視真人秀節目《澤西海岸》中保利·D（Pauly D）和邁克·索倫蒂諾（Mike Sorrentino）的人生哲學。——譯者注

[11] 「駭人」長髮辮是牙買加黑人、雷蓋樂樂師等的一種編成緊緊的辮子的髮式，很多人認為這是不衛生的。——譯者注

[12] 僅僅在2013年12月，該網站就迎來了1.01億位用戶的訪問，這些用戶創造了50億頁的內容。

[13] See 「I'm Just Gonna Throw This Out There.Any Redditors in the SF Bay Area Have a Empty Spot at Their Table for a Lonely Thanksgiving Orphan?」 posted by user 「MeMyselfOhMy」 on Reddit: reddit.com/r/AskReddit/comments/ebhh1/.

[14] 下面提到的例子摘自2013年1月各個主題的頭版。

[15] 安德森關於民族的很多理念都令人驚訝地適用於網絡社區的討論，這表明其理論具有很強的適應性。安德森將民族描述為一方面具有內在的限制性和主權性，另一方面具有深刻的、平等的友愛。下面這一段話特別適用於網絡社區：「只有當一大群人認識到自己能夠與另外一大群人共同生活的時候——即便是從未謀面的人，生活卻沿著同樣的軌跡前行——這種嶄新的、共時性的民族才可能出現在歷史之中。」這段話摘自安德森所著的《想象的社區》一書（London: Verso,1983, 6, 191—192）。

[16] 我從奧德·霍夫雷特納（Aude Hofleitner）、塔·維羅特·奇拉法德漢納庫爾（Ta Viroth Chiraphadhanakul）以及博格丹·斯泰特（Bogdan State）這三位Facebook的研究人員那裡得到了許可，可以複製他們製作的地圖，並討論他們的研究成果。他們請我把他們的一段話加進來，更加詳細地解釋一下「集體遷徙」這個概念以及他們的研究。下面就是他們的補充說明：在集體遷徙中，一個城市的大批人口作為一個群體遷徙到了另一個城市。更加明確地講，假設有A城市（家鄉）和B城市（現居住地），如果說A城市到B城市的人口流動可以被稱為集體遷徙，那麼就意味著從A城市遷出的人裡面，絕大部分都居住在B城市。就美國而言，無論是外國人移

民美國，還是美國人移民他國，還是美國人在國內不同城市之間的遷徙，都非常頻繁，但都沒有表現出集體遷徙的特徵，因為沒有任何一個城市具有特別大的吸引力，人們遷徙的地點較為分散。這張地圖展示了東南亞地區的大批人從小城鎮和農村集體遷到城市中心地帶的情景。如果想獲取更多信息以及他們完整的研究成果，請參考Facebook數據科學團隊發表的關於集體遷徙的帖子，獲取鏈接為：www.facebook.com/notes/facebookdata.science/coordinated-migration/10151930946453859。當你打開這個鏈接時，就會發現我在引用這些數據時進行了一定的修改，移除了一些標籤，並將選圖集中在較小的範圍，以便於印刷。感謝在Facebook工作的邁克·德弗林幫助我取得了相關資料的使用授權。Facebook數據科學團隊在研究過程中使用的資料都經過了匿名化處理和整合。

第十三章 個人品牌

巴斯啤酒公司艾爾啤酒的三角形標識是英語國家的首個註冊商標，當今，其悠久歷史構成了這一品牌吸引力的關鍵。他們也將之寫在了商標標籤上——英格蘭的首個註冊商標。不過，你有所不知的是，巴斯之所以成為英格蘭的首個註冊商標，只是因為英國商標註冊法案生效的那天早上，登記所門前排起了隊，而巴斯啤酒公司的一名僱員恰巧站在了隊首。^[1]他們成功地利用一次到政府機構辦事的偶然機會，將之發展成了自己的一個口碑——至少現在從那些棕色瓶子裡裝的東西來判斷，這個口碑遠遠超過了產品的實際品質。巴斯就是一個完全建立在品牌宣傳之上的品牌。

在巴斯之前就有很多品牌和商標了，於是英國開始對它們進行監管，可見，商標和形象塑造甚至在工業革命之前就存在了。換句話說，品牌原本就是深入骨髓的東西，由來已久。考古學家從5000年前的沙漠墓穴中發掘出了標著自身品牌的石油和葡萄酒。一個在埃及發現的商品標籤在皇家標誌和畫著一個金色榨油機的象形文字下面寫有「泰赫努最好的石油」。^[2]這就好比一罐百威啤酒在「啤酒之王」的字樣下面寫著「精選最優質啤酒花、稻米和大麥」——儘管品牌宣傳發展到了今天，不過在很多方面，它很可能永遠是一門誕生於青銅器時代的學問，因為它所迎合的情感是永恆的。

儘管聲譽或許是有關品牌的永恆概念，但品牌真正的新領域近來已經開闢出來了——人。1997年，激情演說家和管理顧問湯姆·彼得斯在《快公司》（*Fast Company*）雜誌上發表了一篇題為《你就是品牌》的文章，個人品牌化的時代由此到來。^[3]

他的文章其實更像一種推銷術，它要讀者首先確定自己的「特性—利益模式」，然後不遺餘力地將之推銷給僱主、同事以及全世界，要不然的話……讀此文章時，你可以想象到彼得斯先生就像一隻困於籠子裡的獅子一樣在講臺上來回踱步——他囿於自己即將在你眼前用知識炸彈、專業技能及許多感嘆號所展現出來的那種範式。他所顯現出來的那種篤定，會讓一個不同類型的人把他的電話簿撕成兩半。這篇文章最下面的署名行寫著：「在寫作、演講或思考新經濟方面，湯姆·彼得斯是

世界首屈一指的品牌。」在當時，他不僅是一流的，也是唯一一個自稱就是一個品牌的人，因此是一個借鑑了巴斯在維多利亞時代戰術的「新經濟」的代言人。何樂而不為呢？在做到之前就先假裝如此唄。這篇文章引入了自我推銷會使你直接邁向成功這一思想，而且在今天它仍是營銷課堂上的讀物。

幾年之後，一個名叫彼得·蒙託亞（Peter Montoya）的人寫了一本在這一領域很重要的著作《你就是品牌》（*The Brand Called You*）。在書中，他進一步發展了湯姆·彼得斯的觀點。是的，名字和彼得斯的文章名相同，不過，這本書並非兩人合作寫成的。事實上，真要說的話，兩個人在品牌化專家這一行其實是競爭對手。將無知與厚顏無恥結合起來是天下所有收入不菲的推銷員的天賦所在，而蒙託亞或許正是那個天才巫師。他的這本《你就是品牌》基本上就是按一條長的主線展開的其第一個要點出現在第二頁，是：

1.你與眾不同。與眾不同——使自己在他人眼中獨具一格，不同凡響——這是個人品牌化最重要的一個方面。

蒙託亞的《你就是品牌》自然很暢銷，而蒙託亞也像彼得斯一樣，直到今天，演說事業依然蒸蒸日上。不過，「如果成為你自己的品牌」這個觀點僅僅停留在這個國家的會議大廳和宴會廳，然後像冷咖啡和鬆餅碎屑一樣融入時代精神，那我就不會寫關於它的東西了。

這個觀點會傳播開來，而且傳播得很快。現在你看到，一旦某位公眾人物有失態之舉或失寵了，自然而然便會有一個問題：這將會對他的個人品牌起到什麼樣的影響？彼得斯和蒙託亞都是創新者，我真的認為他們是。我所認識的一些最聰明、最理應成功的人說「我的品牌」這幾個字時理所當然。通過「谷歌圖書」，你可以在已經出版的書中看到這個觀點的誕生及其後來的飛速發展。^[4]

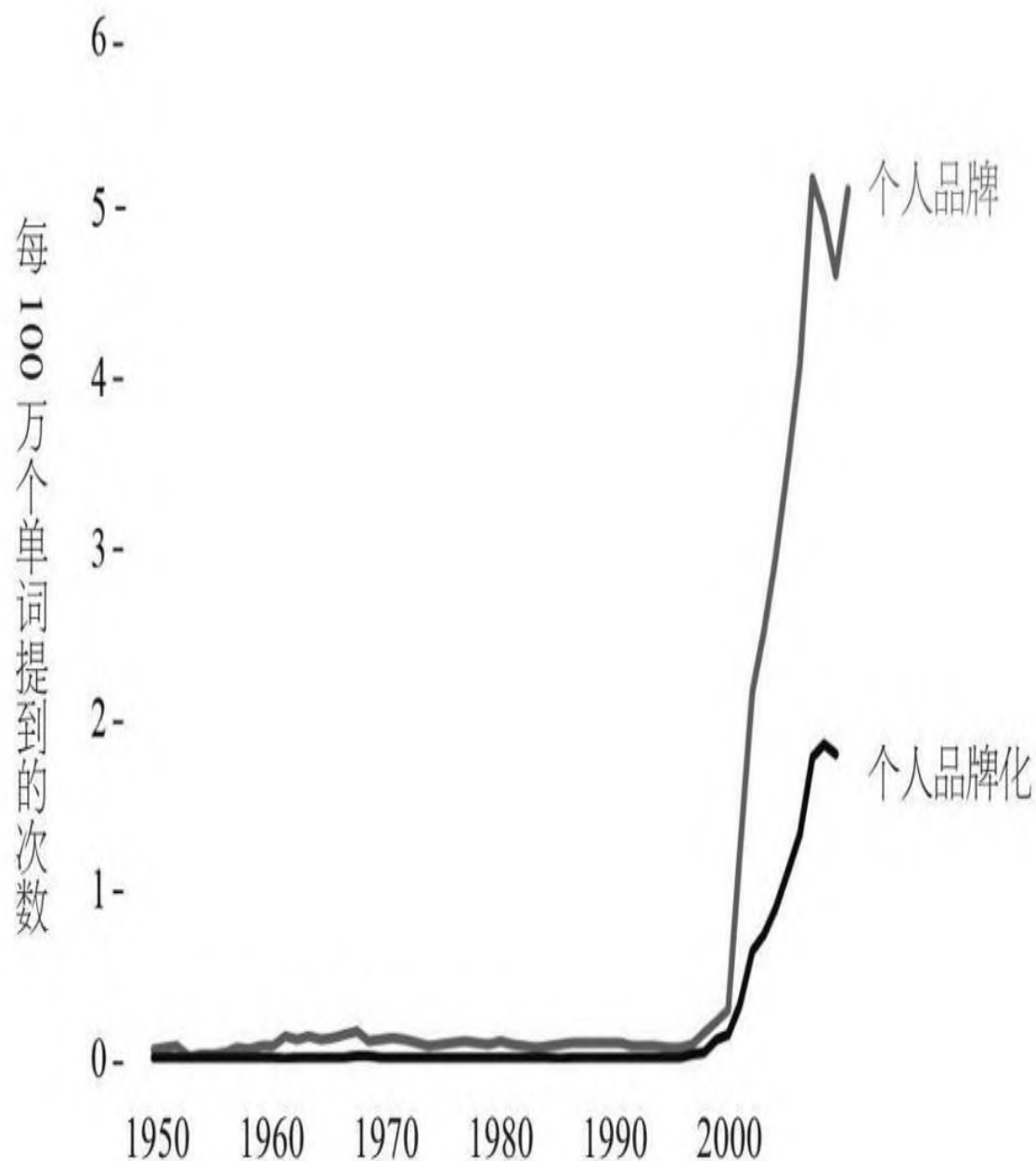


圖13—1 個人品牌與個人品牌化

當然，個人品牌化的原則並非現在才有。彼得斯^[5]和蒙託亞與戴爾·卡內基並非沒有相同之處。^[6]戴爾·卡內基通過借用了鋼鐵大亨安德魯·卡內基的黃金姓氏「卡內基」，重塑了「戴爾·卡內基」這個稀鬆平常的名字，並從中誕生了自己的品牌。他和當今很多人一樣，將個性簡化為幾個要點，並視影響力為使一個人走向成功最關鍵的因素。個人品牌

化的目的始終是財富和權力。

不同之處在於，現在「個人品牌化」要求你通過將自己視作產品而非視作一個人來對待，從而實現這些目的。彼得斯又寫道：

從今天開始，你就是一個品牌。和耐克、可口可樂、百事可樂或美體小鋪一樣，你就是一個品牌。你要開始像自己最喜歡的品牌經理那樣思考，想想耐克、可口可樂、百事可樂或美體小鋪的那些品牌經理會問什麼問題，然後問問自己這個同樣的問題：究竟是什麼使得我的產品或服務不同於其他產品或服務呢？

這是個人品牌化的核心概念，就像基督教加印刷機或職業橄欖球加電視機的組合一樣，在社交媒體中這一觀點找到了可以藉以推廣到全世界的最理想的技術。我不會在此贅述Facebook、Twitter及Instagram等網站是如何讓你將自己投射給這個世界的，不過我要指出，不久以前，只有大公司或者有大預算的公司才能夠使自己的信息為地球另一端的陌生人瞭解並喜愛。現在，我可以，你也可以，每個人都可以。最難的部分在於讓每個人都來聆聽。

最簡單的方法是使自己有意思、吸引人、風趣。不過那些真正能使人們發笑的喜劇演員是很少的，這是有原因的，因為很難。一個業餘演員想要通過在Twitter上表現得詼諧或令人振奮而不斷壯大自己的粉絲群，最後成為下一個賈斯汀·薩科（因為一條帖子而被解僱的公關經理）的可能性要遠遠大於成為下一個賈斯汀·哈爾佩恩（記錄他老爸每日令人叫絕的快言快語而迅速成名的網絡紅人，他的Twitter賬號是ShitMyDadSays）的可能性，後者有300萬粉絲，並且已經簽訂一個出書協議。若有一個孩子靠發Twitter上了大學，或者得到了一個《紐約客》的工作機會——有人做到了——必有幾十個孩子因為發Twitter而被請進了校長辦公室，或者更可能銀鐐入獄。^[7]

運用我們的文本分析算法，你可以看到發展壯大粉絲隊伍所需要的一些東西。表13—1是我稱為「業餘級」及「嶄露頭角的專業級」粉絲水平的一些典型用詞。

表13—1 粉絲數小於100人和超過100人的Twitter用戶典型用語

粉丝数小于 100 的人	粉丝数超过 100 的人
thehungergames (饥饿游戏)	partnering (合伙)
upset (烦恼)	heyboo (喜宝)
worthit (值得)	vamping (拼凑)
whyme (为什么是我)	optimizing (优化)
roethlisberger (罗斯利斯伯格)	sourcing (提供消息)
workaholics (工作狂)	marketer (市场营销人员)
wordsofwisdom (智慧语录)	tweetup Twitter (推特)
hurryup (快点)	visibility (可视性)
depressed (抑郁)	monetize (货币化)
wishmeluck (祝我好运)	industry's (行业的)
getonmylevel (达到我的水平)	optimize (优化)
studying (学习)	brownskin (棕色皮肤)
idiots (傻瓜)	merchants (商人)
cincy (辛辛那提)	influencers (产生影响的人)
collegeproblems (大学问题)	robust (强健的)
sunny (阳光的)	yeen (你不是)
notokay (不好)	guwop (歌手谷沃普)
finalsweek (最后一周)	talmbout (谈谈关于)
tebow (蒂博效应)	innovators (创新者)
silly (傻气)	partnered (成为搭档)
impatient (不耐烦)	bezos (贝佐斯)
leavemealone (别管我)	infographics (信息图)
holysht (天哪)	livest (最活跃的)
suckstosuck (真不爽)	strategist (战略家)
pujols (皮若尔)	entrepreneurial (企业家的)
saveme (救我)	slideshare (幻灯片分享)

(續表)

粉丝数小于 100 的人	粉丝数超过 100 的人
yeahbuddy (是啊, 兄弟)	yass (你的屁股)
pattys (小馅饼)	amplify (放大)
girlproblems (女孩子的问题)	goodmorning (早上好)
killme (杀了我吧)	creatives (有创造力的人)

在表13—1的左列你看到的是那種簡單的、眼前一閃而過的擔憂，從使用Twitter的人的身上你會預料到有這樣的擔憂。在右列，你看到的幾乎完全是管理性的術語：如果你有眾多粉絲，事實上你很可能像一個公司那樣講話。但是右列有些詞並不是典型的專業詞彙，如#heyboo（喜寶）、talmabout（「談談關於」的壓縮形式），yeen（你不是）、yass（你的屁股）等。就像左列的人一樣，他們在Twitter上聊八卦、發牢騷、顯擺，他們只是在更廣泛的圈子裡這樣做，面向幾千個粉絲這樣做。這些詞背後的用戶是黑人，右列的這些詞的存在證明非裔美國人傾向於用不同的方式使用Twitter。（我強調「傾向」是因為沒有哪個群體中每個個體的行為會完全一致。）評論家們稱這一現象為「黑人Twitter」，法哈德·曼約奧（Farhad Manjoo）在《石板》雜誌上有如下描述：

黑人——具體地說，年輕黑人——看起來確實和其他人使用Twitter的方式不同。^[8]他們在網絡上形成了關係更緊密的群體——他們更願意互相關注彼此，他們之間的互相轉發更加頻繁，而且他們的帖子更多是回覆——是直接給其他用戶發的。不管是有意還是無意，正是這種行為，給予了黑人——尤其是黑人青少年——主導Twitter言論的手段。

他說的「主導」指的是：在Twitter剛推出來的時候，白人用戶有些混亂，像「uainthittinitright」和「如果聖誕老人是黑人」這樣的話題標籤會和主持人瑞安·西克萊斯特新出爐的精闢語錄或Old Spice（寶潔旗下男性護理品牌）的獨特新奇的市場營銷手段（就像#heyboomonetize）

一樣，成為Twitter上面的熱門話題（就像heyboo這個話題和上面的「貨幣化」並列出現或許看起來有點令人困惑一樣）。大多數Twitter用戶都會關注某些名人或新聞，只是名人不會也去關注他們。Twitter的主流文化是一對多的交流，事實上是圍繞著品牌進行的。但是黑人用戶往往主要是用作私人用途，並且具有高度的互惠性。因此他們的粉絲很多，而且也更有能力在圖表上使自己的模因凸顯出來。

任何希望用主流的方式在Twitter上建立自己品牌的人——成為那個對眾人發聲的人——應該意識到Twitter其實很大程度上是1%的人的世界。它最寶貴的資源——粉絲的分配要比財富分配不公平得多。在我所收集的樣本中，頂端1%的人的粉絲數佔了總數的72%，而頂端0.1%的人的粉絲只佔0.5%。擁有100萬粉絲比賺100萬美元要難得多。2011年向美國國稅局報告自己收入超過100萬美元的人有30.089萬。現在，在全世界範圍內，擁有100萬粉絲的Twitter賬號有2643個，或許其中有一半是在美國。^[9]在美國，在Twitter上擁有100萬粉絲大約相當於是個擁有10億美元的富翁。^[10]

當然，前提是所有粉絲都是真實的。我為自己的某個賬號買了一些粉絲，以便看看它是如何運作的。比如在一個叫TwitterWind的網站上，你可以從菜單上挑一個數字（我選擇的是1 000），並付錢，一兩天之後，幾乎是一瞬間，你得到了許多沒有什麼用處的新朋友。這些僱來的粉絲什麼都不做，只是存在而已，而幾乎每個Twitter粉絲數量眾多的人很可能都買了一些——尤其是那些一心想出名的人，比如名人和政客。在共和黨總統候選人提名之戰懸而未決的時候，紐特·金裡奇就放言：「我的Twitter粉絲是其他候選人數量加起來的6倍。」^[11]他唯一未透露的是其中90%的粉絲都是他花錢買來的。^[12]米特·羅姆尼（幾乎肯定）也買了粉絲，比如，他在7月的某一天裡，粉絲數量在短短的幾分鐘內就增加了2萬，而這是他在這之前和之後相同時間裡粉絲增加數量的200倍。^[13]現在，請注意兩個要點。一是，一個人可以為其他人買粉絲，那麼這很可能是某個新世紀的尼克松在玩政治間諜魔法。這當然是個讓羅姆尼看起來像個蠢貨的好辦法。二是，我敢肯定，奧巴馬和許許多多民主黨人也為自己買過粉絲。為了操縱制度為己所用，就採取厚顏無恥的行為是兩黨的家常便飯，只是他們不像圖13—2所顯示的這樣容易被發現。

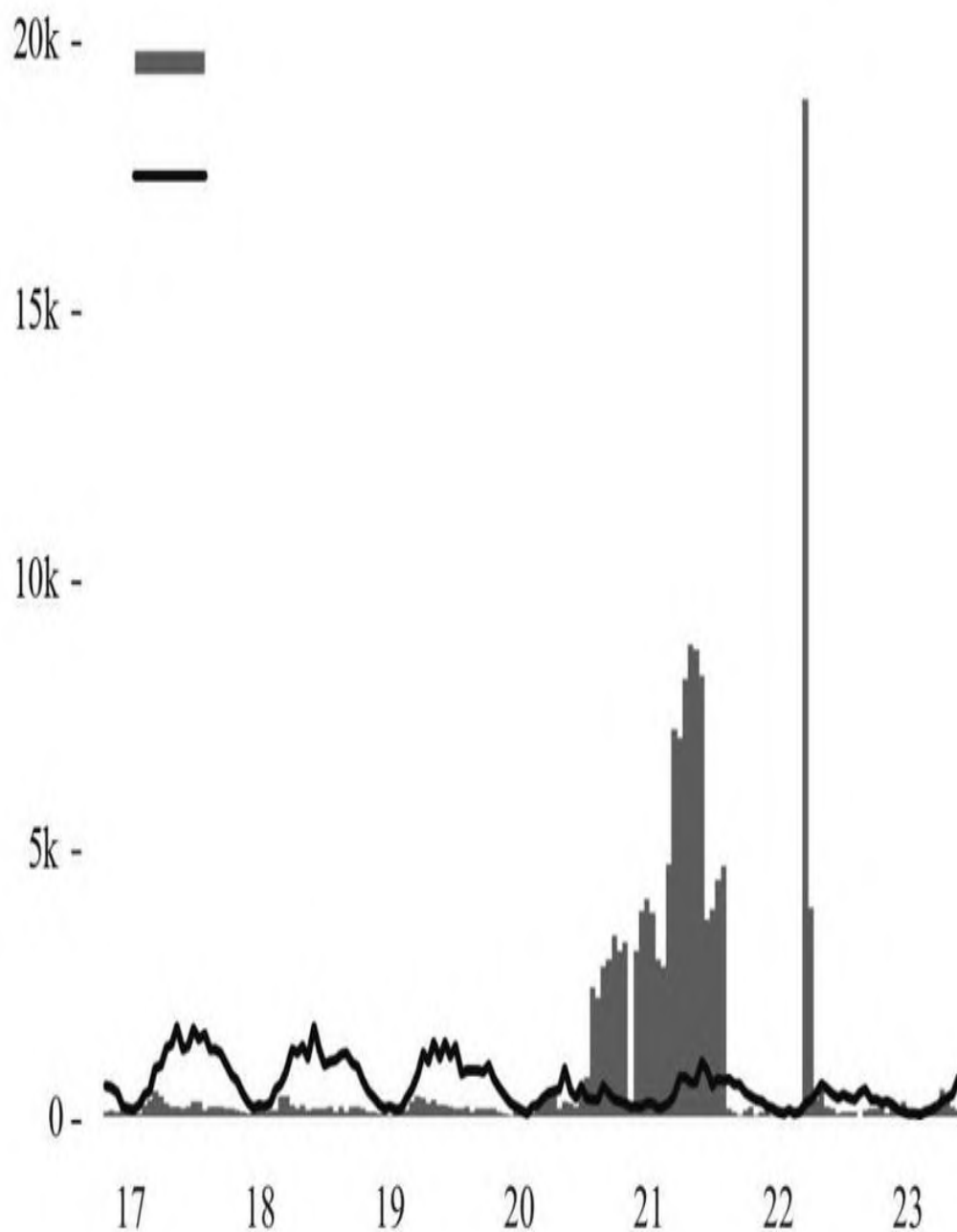


圖13-2 米特·羅姆尼的粉絲變化（2012年7月17日—23日）

你能理解為什麼這些人這麼做。一個人看起來人氣越高，他們的實際人氣就會越高。普通的Twitter用戶也同樣難以抵擋這種輕而易舉就能得到朋友的誘惑，儘管他們沒有奧巴馬和羅姆尼在這方面的預算多。在我隨機抽取的Twitter數據集群中（數字1和數字5），最常見的5個話題標籤中的兩個分別是#ff和#teamfollowback（回跟團隊）。#ff代表的是

「關注星期五」，這是Twitter上的一箇舊傳統。現在它是「嘿，關注這些賬號」的通用簡稱，通常會從那些想要粉絲數量增加的用戶嘴裡喊出。「#回跟團隊」是一個Twitter賬號的標籤（手柄）——它的功能是不用花錢就能做到政客們花錢做的事。這個理念是如果你關注「回跟團隊」，這個賬號的其他粉絲就會來關注你，然後，你也反過來再去關注他們。這樣，每個人的粉絲數都漲了。這和同盟網站的舊理念一樣——同盟網站在谷歌出現以前是各個網站相互點擊、確保訪問量的一種方法。下面是「回跟團隊」的自我描述：

我們保證你會得到回跟的粉絲！這是我們原創的，也是最棒的——推廣我們的標籤#WILLFOLLOWBACK（會回跟）#TEAMFOLLOWBACK（回跟團隊）吧

「自己就是品牌」這一思想導致的結果就是：追逐虛無的標準。我知道自己發Twitter時，我對誰分享了、分享得有多快的關注度不亞於我對自己本來想要表達的內容的關注。我不怎麼在Facebook上髮狀態，可每次發時，我都會坐在那裡，不斷刷新頁面，以看到新的評論，就好像我是剛剛開始使用互聯網一樣。《紐約時報》的珍娜—沃瑟姆把這種心態描述得很好：「我們——用戶、生產者、消費者——近乎瘋狂，迫切希望自己對正在進行的討論所做的重要貢獻被人關注、認可——這是問題所在。而這一問題又因為我們不斷需要通過喜歡、最喜歡、迴應、互動、被人關注、得到肯定而愈演愈烈。它是一個無法結束的反饋迴路，至少目前無法結束。」^[14]我可以告訴你內幕：公司設計產品就是為了讓這個反饋迴路永不結束。OkCupid會給你看你的消息、訪客、可能性的數量。我們知道是這些數字使我們的用戶一直對它感興趣，當這些數字增加的時候更是如此。如果沒有那點小小的興奮，網頁或者應用就無人問津，而用戶也會漸漸離開。這一現象的一個廣義的術語是「用戶參與度」，即有多少人每週、每天、每小時都在登錄查看。網站顯示你的各類數目、總數、徽章，因為他們知道你會回來看這些數目增加了多少，然後他們就可以把你不斷提升的參與度放在幻燈片上以給投資者留下深刻印象。

這就是問題所在：你把自己變成數字是一回事（要另當別論）；而如果有人把你變成數字，那感覺會很糟糕。Klout是一家頂級個人分析公司，他們看著你所有的社交媒體賬號，然後通過自己的一個小小的獨家魔法，全面衡量你的在線影響力，給出一個分數（從0分到100分）。蒙託亞（及卡內基）的例子會讓你銘記：影響力是你個人品牌的全部。

Klout會幫你搞清楚你的影響力到底有多大。現在，我的Klout分數相當可憐，只有34分。回跟團隊的分數為60分，這使我既想哭又想笑。一方面，他們存在的原因只有一個，而他們相當於得到了D的成績；另一方面，他們比我認識的任何人的分數都要高。

2012年，雲計算巨頭Salesforce.com網站發佈了一則招聘公告，上面列的一項「理想技能」是Klout分數不低於35分。^[15]這一項並非必要條件，不過該網站將之與其他明顯重要的特質，比如「團隊合作能力」放在了一起，所以想必這一項是這個工作的核心部分。Salesforce.com的業務主要是進行量化——他們幫助公司通過數據進行市場推廣，^[16]所以，他們以同樣的量化方式來招聘員工也就不足為奇了。但是即使像信用分數這樣的數字，有段時間也成了招聘過程中令人討厭的一個部分，看到Klout分數被列為招聘條件還是會使很多人不高興。

BetaBeat網站上的一篇文章《想來Salesforce工作嗎？最好要讓自己的Klout分數至少達到35分》招來的評論「呸」就準確道出了公眾對此的普遍反應。然而，真正的擔心是：我們要淪為數字了，而且，很快，人們會在很長時間裡都談論這個問題。Salesforce過去是，現在也還是開啟潮流的公司——至少在在線市場營銷領域如此。他們發佈那則招聘廣告的同一年，《福布斯》雜誌將其列為「美國最具創新性的公司」。他們每年招聘幾百名新員工，更加切題的是，當獲得殊榮的創新公司採取新的做法時，其他公司會效仿他們的做法。如果現在Salesforce對Klout分數有要求，那麼其他公司也很快就會對僱員的Klout分數有要求了。人們不想淪為一個由公司製造出來的兩位數數字，而這家公司即使在虛無縹緲的新興社交媒體世界也似乎有點不靠譜。

但是，考慮到Klout使用了很多我收集數據時使用過的工具，這讓你、我以及我們一直以來傾注心血的這本書怎麼辦？嗯，簡短的回答是：就跟Klout和Salesforce一樣吧。淪為數字是不可避免的。算法是赤裸裸的，計算機是機器啊。數據科學正試圖使一個模擬的世界數字化。它是微型集成電路片的基本物理性質的副產品：芯片不過是一些微小門電路的排列，不在於互聯網就是一系列的電子管，而是事實如此。門電路時開時合，以讓電子通過。^[17]門電路的狀態非此即彼——電路要麼開著，要麼閉著；沒有或許這回事。一種絕對主義從顯微鏡下的現實中誕生，並在整個行業中滋生壯大，直到在最高層面，你有了定義、數據類型、類別等編程語言（如C語言和JavaScript）的核心內容。

就這樣，信息因客觀需要而被簡化。但是從根本上來說，對Klout

分數要求的抵制關乎人們（不僅僅是他們的信息）被簡化成數字。下文指出了本書與Salesforce的工作職位要求以及事實上整個Klout的業務模式的不同之處。

儘管有很多數字，但是數字不能評價任何人。單一數字從來都不能評價一個人。有個未必真實的故事是說，愛因斯坦在高中時曾數學不及格，這個故事就道出了這個真理。他數學不及格這種可能性是存在的，而且如果他真的數學不及格，誰又會在乎呢？如果他幾何II得了35分，那又如何呢？難道他突然間就變得不聰明瞭嗎？任何數字、考試、單一衡量標準——包括智商、身高，當然還有Klout分數、OkCupid上的朋友數量或者回覆比例——都不能說明一個人，這也正是為什麼除了圖表之外，任何個人用戶都不曾出現在我的書中的原因。但是，如果把關於我們的一些細節的側面信息集合起來，就得到了一些宏觀的東西。大量數字的法則是我們以前多次忽略的一個觀點，不過我想明確說一下：數據的全部真相只能通過一個大的樣本來揭示。假設有一個神奇的色子——你不能數它的邊，但是你能滾動它，看看得到的數字是幾。滾動一次，你可以得到一個數字，不過這時你什麼規律都找不到。滾動幾次，你可以知道數字分佈情況，你得到了平均數——這就完全說明了色子的特性。你只能通過大量情況的彙集來知道它的形狀。

而且，簡化和重複在科學的漫長發展史中起著根本性的作用。實驗的基礎是將一個過程簡化成單一、可控的側面。科學方法需要一個參照標準，你若不將其複雜性簡化成最基本的核心所在，並承認這就是重要的東西，你就不能得到它。一旦你簡化了這個問題，你就能一次又一次地對它進行檢驗了。不管是在實驗室的工作臺上還是筆記本電腦上，我們所擁有的大多數知識都是這樣獲得的——通過簡化。

所以，我們已經將人性歸結成了數字，而非故事。在我看來，本書毫不非主流。我不是要偏離數據所能觸及的巨大範圍——絕無僅有的東西、各種例外和單一情況，以及那些你需要知道所有情況才能得出真相的事例，比如上面提到的愛因斯坦的故事——我正在從那些沒有分別的整體中發現問題。我們關注的是密集的集群、中心所在，以及由於人類經歷的重複性和共通性而一再被複制的數據。這就像點描式的科學。那些點或許只代表我們的一小部分，但是全部的點卻構成了我們。

集聚和重複也使我們得以把握大趨勢，大趨勢的平穩推進或許不像通常的英雄敘事那樣跌宕起伏，但是這也使它更具適用性。保羅·麥卡特尼和約翰·列儂練習了1萬個小時的搖滾音樂，然後有了甲殼蟲樂隊。

這的確說明了反覆操練和堅持不懈的價值，但是那個數字1萬本身並沒有任何意義。我自己也和許許多多的無名音樂家一樣，花了那麼多時間彈吉他。使得列儂和麥卡特尼把練習變成天分的東西是他們所獨有的。另一方面，本書中的每一個數字，背後都有好幾百，甚至好幾千的人，他們都不為人知。其實質在於：「百萬挑一」是許多令人歎為觀止的藝術作品的核心所在。它意味著一個人是如此特別，如此有才華，如此難得，他們事實上是獨一無二的，正是這種稀有使得他們非比尋常。但是在數學裡，和數據一樣，和在這本書裡一樣，「百萬挑一」的意思正好相反：一百萬分之一是舍入誤差。

但是如果理解大的數據集群需要進行簡化，我真的要為一種不同的簡化法而擔心了，那就是，人們不是僅僅淪為一個數字，而是淪為一個失去人性的用戶碼，要被塞進一個市場營銷的算法裡，成全了某個其他人的品牌。數據使推銷術去除了許多猜測的成分。它是少有的一個最終證明正確的都市神話，但是塔吉特公司通過分析一位顧客所購買的東西，得知她已經懷孕了，而她還未告訴任何人這件事。^[18]問題在於，她還是個少女，所以他們開始把與母嬰相關的廣告寄到她父親家裡去。

在某些方面，這種公司對個人生活的介入比品牌試圖與你「建立聯繫」要好。前幾年的一個夏天，一個Jell-O（吉露果子凍）的市場營銷活動選用了#fml這個話題標籤——在互聯網上，fml是「糟糕的人生」

（fuck my life）的簡寫。^[19]他們負責社交媒體的工作人員開始對帶有這個標籤的Twitter進行迴應，並主動提出要使參與者的人生變得有趣，而不是贈送優惠券。這樣處於低谷中的人們得到了該公司滿心歡喜提供的東西。下面就是一個例子：

Pyrrhus Nelson @suhrryp Seeing my bank account disappear at the dr office #fml

（我的錢全砸給醫生了。#糟糕的人生）

JELL-O @JELLO @suhrryp Fun My Life? Of course we will.In fact, we'd be happy to.prmtns.co/dkTq Exp.48hrs

（想要生活充滿樂趣嗎？當然想！其實，我們一直這樣想。）

這種其他用戶並不想要的介入在社交媒體上太容易了，因為一切都是量化的。話題標籤直接就跳到了品牌經理的屏幕上，他通過提供折扣參與了進來。不過，這使得他們進入我們生活的同樣技術也使我們能夠反擊。幾年以前，麥當勞發了幾條Twitter，是關於他們的供應商的正面

故事，話題標籤是「麥當勞故事」，相反他們得到了#fml。^[20]下是眾多回應中的一個：

MUZZAFUZZA @Muzzafuzza

I haven't been to McDonalds in years, because I'd rather eat my own diarrhea.#McDStories

（我已經好幾年沒去過麥當勞了，因為我寧願吃自己的便便，也不想去那裡。）

麥當勞付了錢以推廣這個話題標籤，但是在僅僅幾個小時之後就撤回了這項活動，因為活動很快就變得無法控制了。一週以後，這個已經改頭換面的「麥當勞故事」標籤依然勢頭強勁。他們的社交媒體策略師本該知道事情會怎樣發展：幾個月之前，溫迪國際快餐連鎖集團本打算推廣「這是牛肉」這個標籤，結果他們的口號完全脫離了他們本來設想的語境。^[21]人們用它來對任何他們不喜歡的東西進行抱怨，而忽略了這個品牌本身，例如：

RemiMitchison @RemiBee

#HeresTheBeef when a chick see another chick doin better and has more than she does ...so she wanna stunt and #GetThatAssBeatUp

（當一個人看到另一個人做得更好、擁有的東西更多時，她就會想辦法掌握絕技超越對方。）

Jeremy Baumhower @jeremytheproduc #HeresTheBeef The drugs companies have already cured HIV and cancer, however it is far more profitable to keep people barely alive on drugs

（藥企或許已經能夠治癒艾滋病和癌症，但讓人們天天吃藥更有利可圖。）

最近一段時間，百事公司下屬的激浪品牌舉行了一場「給激浪起個新名字」的比賽，想要乘著「集體智慧」的東風，給自己的飲料起個炫酷的新名字，心想或許如果一切順料的話，這些評估標準就會顯示出足夠的動力讓真正有影響的人買入，自己就能在博客圈贏得一些品牌大使。^[22] Reddit和4chan這兩個網站捕捉到了這個機會。結果，在網民的留言中，「希特勒沒做錯什麼」領先了一段時間，後來突然出現了「喝這種飲料會引起糖尿病」的信息，最後變成了：「取什麼名啊，渾蛋！」

互聯網是個瘋狂的地方，不過也正是它可能帶來一些不可思議甚至瘋狂的東西。我想象不出來為什麼這個品牌會導致網友想到為希特勒翻案！此外，這何等浪費時間！因為激浪公司並不打算在其寶貴而獨特的標誌上印上一個不利於自身形象的詞語。大張旗鼓地為蘇打飲料徵名並不是一個值得推薦的做法。我在這種糊塗、輕浮甚至愚蠢中找到了安慰。這種做法對於任何行業來說都不明智。這也證明，儘管社團主義（corporatism）或許會侵入我們的新聞提要中，甚至就像一些人所希望的那樣侵入我們的靈魂，但我們身上仍有一小部分是它無法觸及的。這就是我始終想要銘記的：使我們喪失人性的不是數字，而是我們深思熟慮之後決定不再把自己當作人。

[1] See Clare Baker, 「Behind the Red Triangle: The Bass Pale Ale Brand and Logo」 Logoworks.com, November 8, 2013, logoworks.com/blog/bass-pale-ale-brand-andlogo/.

[2] My discussion of branding in ancient times is based on David Wengrow, 「Prehistories of Commodity Branding,」 *Current Anthropology* 49, no.1 (2008): 7—34, and Gary Richardson, 「Brand Names Before the Industrial Revolution,」 NBER Working Paper No.13930, National Bureau of Economic Research, Cambridge, MA, 2008.[http:// papers.nber.org/paper/w10411.3](http://papers.nber.org/paper/w10411.3). See 「The Brand Called You」 by Tom Peters, published in *Fast Company*, August/September 1997, fastcompany.com/28905/brand-called-you.

[3] 關於這一點，請參考一位名叫「Morgan」的網友在Fastcompany.com網站上對比特的文章的評論，該評論認為：「這篇文章真是太好了。這是我們上營銷課必讀的文章，寫得很好，見解深刻，信息量大，謝謝！」

[4] 在圖13—1中，我將personal brand of（某個人認為自己是……的代表）這個表達方式從personal brand（個人品牌）裡剔除了出去，比如personal brand of leadership（個人認為自己是領導力的代表）就排除在考慮範圍之外，這樣就可以確保我呈現出來的person brand都是「個人品牌」的意思。

[5] 彼得斯的一個口頭禪就是：「要麼與眾不同，要麼與世長辭。」

[6] 關於卡內基的個人資料，我參考了維基百科「Dale Carnegie」條目的內容。

[7] The two incidents I allude to here are Bernie Zak's campaign to get into UCLA, as detailed in Brock Parker, 「Brookline Student Lobbies UCLA on Twitter」 *Boston Globe*, May 7, 2013, and Rob Meyer's hiring by the *Atlantic Monthly*, as described in Alexis C. Madrigal, 「How to Actually Get a Job on Twitter,」 *Atlantic Monthly*, July 31, 2013. See also Jason Fagone, 「The Construction of a Twitter Aesthetic,」 *The New Yorker*, February 12, 2014, newyorker.com/online/blogs/culture/2014/02/the-construction-of-a-twitter-aesthetic.html.

[8] 我在探討「黑人Twitter」現象時，參考了下面的資料：ChoireSicha, 「What Were Black People Talking About on Twitter Last Night?」 *The Awl*, November 11, 2009, theawl.com/2009/11/what-were-black-people-talking-about-on-twitter-last-night. Farhad Manjoo, 「How Black People Use Twitter,」 *Slate*, August 10, 2010, slate.com/articles/technology/technology/2010/08/how_black_people_use_twitter.html. A counterpoint to Manjoo's piece is 「Why They Don't Understand What Black People Do on Twitter」 by Dr. Goddess, on blogspot.Goddess especially objects to the portrayal of blacks on Twitter as a 「monolith」——the word appears twice in the post, and I echo it in my discussion. See

drgoddess.blogspot.com/201% 8/ why- they dont understand- what black.html. 「How to Be Black Online,」 a slideshow by Baratunde Thurston, is a clever overview of Black Twitter and acknowledges better than most sources that, like many racial tropes, 「Black Twitter」 is both funny because it's true and inaccurate at the same time. See slideshare.net/baratunde/ how- to be black- online by- baratunde. Hard data on Twitter usage by ethnicity can be found in the Pew Research report 「Demographics of Key Social Networking Platforms」 (2013), by Maeve Duggan and Aaron Smith: pewinternet.org/2013/12/30/demographics-of-key-social-networkingplatforms/. For evidence of white confusion over Black Twitter, see Nick Douglas, 「Micah's 'Black People on Twitter」 Theory, Too Much Nick, August 21, 2009, toomuchnick.com/post/168222309/.

[9] Social Bakers 這個網站依據粉絲數量對所有 Twitter 賬戶進行排序。當然，如果你現在登錄這個網站去看的話，排序情況肯定與我寫這本書時查到的情況存在變動。For information on US taxpayers by income, visit the IRS's 「SOI Tax Stats—Individual Statistical Tables by Filing Status」 page at irs.gov/uac/2013-12-17-soi-tax-stats-individual-statistical-tables-by-filing-status. Information on the Forbes Billionaires list is from Elizabeth Barber, 「Forbes' Richest People: Number of Billionaires up Significantly,」 Christian Science Monitor, March 3, 2014, csmonitor.com/USA/USA-Update/2014/0303/Forbes-richest-people-number-of-billionaires-up-significantly-video.

[10] 2014年，《福布斯》富豪榜上財產超過10億美元的富豪共有1645人。

[11] See Jeff Neumann, 「Newt Gingrich Brags About His Twitter Followers,」 Gawker, August 1, 2011, gawker.com/5826477/newt-gingrich-brags-about-his-twitter-followers. Also see John Cook, 「Update: Only 92% of Newt Gingrich's Twitter Followers Are Fake,」 Gawker, August 2, 2011, gawker.com/5826960/newt-gingrich-twitter-followers-are-fake.

[12] 金裡奇的一位前幕僚曾經對Gawker博客說：「這些賬戶裡，大約80%是不活動的或者說是虛假的賬戶，是由各種專門幫人增加粉絲的公司創建的。另外10%的賬戶是跟金裡奇互粉的人，他們也會出錢購買粉絲。還有最後10%的粉絲是真正喜歡金裡奇的人。」

[13] See 「Is Mitt Romney Buying Twitter Followers?」 by Zach Green on 140elect: 140elect.com/twitter-politics/is-mitt-romney-buying-twitter-followers/. My data and chart are adapted from the data and chart in that post.

[14] See Jenna Wortham, 「Valley of the Blahs: How Justin Bieber's Troubles Exposed Twitter's Achilles' Heel,」 New York Times Bits blog, January 25, 2014, bits.blogs.nytimes.com/2014/01/25/valley-of-the-blahs-how-justinbiebers-downfall-exposed-twitters-achilles-heel/.

[15] 當我寫到這一段時，這個招聘廣告還在，但現在已經移除了。我對該公司招聘情況的討論參考了下面兩篇文章：Drew Olanoff, 「Klout Would Like Potential Employers to Consider Your Score Before Hiring You. And That's Stupid,」 TechCrunch, September 29, 2012, techcrunch.com/2012/09/29/klout-would-like-potential-employers-to-consider-your-score-before-hiring-you-and-thats-stupid/. Jessica Roy, 「Want to Work at Salesforce? Better Have a Klout Score of 35 or Higher,」 BetaBeat, September 27, 2012, betabeat.com/2012/09/you-may-not-work-at-salesforce-unless-you-have-a-klout-score-of-35-or-higher/.

[16] 作為一家名副其實的分析公司，他們甚至還擁有數據公司。

[17] 門電路不同於需要通過鎖鏈控制開關的門，而是通過電壓來控制門電路的開與關，從而控制電子在不同空間的移動。具體請參考：See Larry Wissel, 「How Does a Logic Gate in a Microchip Work? A Gate Seems Like a Device That Must Swing Open and Closed, Yet Microchips Are Etched onto Silicon Wafers That Have No Moving Parts. So How Can the Gate Open and Close?」 Scientific American, 「Ask the Experts,」 October 21, 1999, scientificamerican.com/article/howdoes-a-logic-gate-in/.

[18] See Kashmir Hill, 「How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did,」 Forbes, February 16, 2012, forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/.

[19] The Jell-O discussion and illustrative tweets are drawn from Harry Bradford, 「Jell-O's Fun My Life Twitter Campaign: Social Media Genius or Just 'Fun'ning Annoying?」 Huffington Post, May 24, 2013, huffingtonpost.com/2013/05/24/jello-fun-my-life-twitter_n_3332230.html.

[20] Drawn from Hannah Roberts, 「#McFail! McDonalds' Twitter Promotion Backfires as Users Hijack #McdstoriesHashtag to Share Fast Food Horror Stories,」 Daily Mail, January 24, 2012, dailymail.co.uk/news/article-2090862/.

[21] Drawn from 「When Twitter Hashtag Promotion Marketing Goes Bad #HeresTheBeef」 by blogger 「stacie,」 on the Divine MissMommy blog: thedivinemissmommy.com/when-twitter-hashtag-promotion-marketing-goes-bad-heres-the-beef/.

[22] See Everett Rosenfeld, 「Mountain Dew's 'Dub the Dew' Online Poll Goes Horribly Wrong,」 Time, August 14, 2012, newsfeed.time.com/2012/08/14/mountaindew-dub-the-dew-online-poll-goes-horribly-wrong/.

第十四章 蛛絲馬跡

Facebook在2009年發佈了點贊按鈕，改變了人們共享內容的方式。但這個理念並不新鮮，因為一些曾經頗受歡迎但現在邊緣化的網站（比如digg.com和del.icio.us）都曾為用戶提供過類似的機會，只是用戶並非直接點贊，而是用五角星對文章進行評分，給予的五角星越多，表示評分越高。Facebook只是對原本就已蓬勃發展的社交網絡進行了內容管理方式上的變革，只要點擊一下那個豎起大拇指的小圖案就可以了，從而更加便於用戶點贊。Facebook創造了一種新的「微型貨幣」

（microcurrency）——我欣賞了你的文章、音樂或其他內容之後，雖然可能不會給予金錢回報，但我會用點贊來報答你，將你的內容分享給我的朋友們。從2013年5月開始，Facebook上每天都會至少點贊45億次，截至當年9月，該網站上共點贊1.13萬億次。^[1]

就在Facebook發佈點贊按鈕的那一年，麻省理工學院的幾名學生髮明瞭「同性戀雷達」，他們的算法擅長猜測一個人的性取向。其工作原理也是非常明顯的，因為男同性戀者的朋友很有可能也是男同性戀者，這不是什麼大祕密，不過其創新之處就在於利用宏觀數據去做微觀事情。自那之後，預測類軟件的能力取得了迅速進步：隨著可用的數據越來越多，這類軟件變得更加智能，運行速度變得更快。2012年，英國的一個研究小組發現，僅僅從一個人點讚的情況，就能判斷出下面這些特徵，只是準確率有所不同（見表14—1）。^[2]

表14—1 從一個人點讚的情況判斷其相關特徵

特征	准确率
男同性恋或男异性恋	88%
女同性恋或女异性恋	75%
白人 or 非洲裔美国人	95%
男性 or 女性	93%
民主党人或共和党人	85%
吸毒或不吸毒	65%
父母是否在其 21 岁之前离异	60%

在這裡，我要再次強調一下，判斷上述特徵的依據並非看一個人的狀態更新、評論、分享或鍵入的其他內容，而是僅僅看點贊情況。你知道科學的觸角逐漸伸到了未知的領域，可能其他人通過點擊幾下鼠標，就能聽到你父母吵架的聲音。一個人的點贊情況甚至可以透露出其智商，可以較為準確地預測出其在標準智商測驗中的得分情況。換言之，不用直接回答問題，僅僅看其點贊情況就能大致猜出其智商水平。

為了得到上述數據，我用三年時間收集了大量用戶的數據，但 Facebook 畢竟是 2004 年 2 月才誕生的，這些用戶在十幾年甚至幾十年的生活中沒有用過 Facebook。如果一個人從兒童時期就開始使用 Facebook 的服務，那麼可能會出現什麼情況呢？如果能收集到這類數據，我會非常激動。但由於該網站上線時間只有 10 年左右，我們現在還無法收集到這類數據，這也是縱向數據的一個不利之處。長期以來，僱主、學校和軍方都用邁爾斯—布里格斯測試（Myers-Briggs）和斯坦福—比奈測試（Stanford-Binet）來測試一個人的智商。他們讓你坐下來，測出最佳狀

態，然後根據你的分數進行分類。大部分情況下，你會選擇參加這類測試。但人們日益清楚地認識到，在測試過程中，人們是根據自己的生活經歷來做選擇的，測試結果是供別人閱覽和評判的，在求職面試之前，一個人的Klout是51分是一回事，但其真實智商是另一回事。

如果僱主用數學算法系統來推斷你的智商或是否吸毒，那麼你唯一的選擇就是同這個系統展開博弈，或者借用前一章的話，你要「管理你的品牌」。要打敗這臺機器，首先你必須表現得像一臺機器，這意味著你已經屈服於機器。要在數學算法中得到高分，首先你要能猜測一下自己應該做什麼。在這類測驗中，一個與測驗智商有關的選項是問受試者是否喜歡「捲曲薯條」。如果你要接受這類測驗，只能回答問題。誰能重新設計問題呢？

儘管Facebook的確知道很多關於你的事情，但對你而言，它更像一位「工作上的朋友」，雖然你們在一起的時間很多，但你們的關係仍然存在明顯的界限。Facebook僅僅知道你登錄之後所做的事情，卻不瞭解你在其他許多場合所做的事情。如果你有一部蘋果手機，蘋果公司就可能知道你的通訊錄、日程表、照片和短信，可能會知道你聽的音樂和你去的地方，甚至知道你走了多少步才到那個地方，因為手機有一個內置的陀螺儀。沒有蘋果手機？沒關係，你的生活中肯定離不開谷歌、三星或威瑞森無線通信。你會佩戴能量腕帶（FuelBand）嗎？如果佩戴，那麼耐克公司就會知道你的睡眠質量。你會使用微軟的Xbox One遊戲機嗎？^[3]如果會，那麼微軟公司就會知道你的心率。^[4]你使用信用卡嗎？如果用，那麼你在一個零售店購物時，你的個人認證信息就會立即將通用產品代碼同你在客戶關係管理軟件中的客戶身份聯繫到一起。

這只是公司數據狀態的冰山一角，如果要進行完整描述，可能需要花上好幾頁的篇幅。對於政府掌握了多少關於公民的數據，我只能說出一丁點兒，因為我只知道這麼多，這些數據是我們能夠公開查詢到的。比如，我們都知道英國全景的閉路攝像頭多達590萬個，平均每11個英國人就擁有一個監視攝像頭，而其中的75萬個設置在醫院、學校等敏感地點。^[5]1998年，紐約曼哈頓區的街道上共發現2397處視頻監控攝像頭，到2005年，僅第十四街以南的地區就發現4176處攝像監控。^[6]衛星和無人機完成馬路以外的圖片。雖然我無法準確地告訴你政府看到了什麼，但我可以肯定地說，如果政府對你的行蹤感興趣，一定會有人盯著你。此外，正如愛德華·斯諾登透露的那樣，有很多事情政府無法通過攝像頭去監控，但可以通過美國國家安全局的內部網絡（NSANet）終

端輕鬆地實施監控，但他沒有透露這個終端的位置在哪兒。

因為很多事情都是在公眾不知情的情況下發生的，所以，公眾肯定無法瞭解到真實的數據。我可以肯定地說，就在我寫這本書期間，已經發生了很多數據，我肯定有些落伍了。從很多方面來講，分析方法比信息本身更重要。網頁瀏覽器裡的cookie^[7]洩露用戶隱私以及黑客竊取信用卡信息的事情時常見諸報端，這類數據蒐集軟件或黑客是最令人苦惱的。黑客可能只是控制了你一小部分生活，為了做到這一點，他們投入了大量的精力。無論他們的腳本語言多麼精妙，他們都是「惡人」，就像很久之前無聲電影中那些長著蓬鬆的大鬍子、戴著高的大禮帽的壞人一樣。或者借用一個較為現代的形象就是《王牌大賤諜》裡的「邪惡博士」。他處心積慮地在月球裝設了能量超強的激光炮，座標對準美國華盛頓，除非美國政府繳納100萬美元的贖金，否則就毀滅地球。他費盡心血，只是為了勒索這一點點贖金，而阿克誠（Acxiom）之類的大數據管理公司卻能輕輕鬆鬆地賺數十億美元。這些大數據管理公司能獲取人們的銀行與信用卡交易記錄及納稅記錄等政府文件，因此能準確地推測出人們的行為方式。那些試圖通過網站調查得出結論的學者無法研究得這麼深入和精準。^[8]與此同時，國家安全機構使用的資源和技術多而複雜，以至企業發明的數據挖掘軟件看起來就像掃雷一樣。

雖然我們在這裡採用了「挖掘」這個比喻性的說法，但這些數據不是一種天然資源，它產生於某個地方，而這個地方就是你。目前，很多公司和政府正在收集與民眾個人生活有關的細節，以便更好地理解民眾的行為方式。你失去的隱私越多，可能影響就越嚴重。在關於隱私問題的討論中，一個最根本的問題就是權衡問題，即你失去隱私之後能得到什麼。我們每時每刻都在做這種權衡。公眾人物失去個人生活，而他們的職業生涯得到了更好的發展。如果一個人為了保護隱私而到歐洲去預訂房間或到印度買火車票，那麼他就必須權衡一下為了隱私空間而多花了那麼多錢是否值得。很多人，既包括男性也包括女性，晚上出門時，為了吸引他人的關注，往往會穿較短或較為緊身的衣服，從而犧牲了自己的隱私。由此可見，這種權衡並不新鮮，但數據的收集和應用方式卻是新鮮的。如果你是一位經濟學家，那麼你會對大數據管理公司出售的分析結果非常感興趣，因為從理論上來講，這些數據意味著廣告更有針對性，意味著營銷費用會減少，意味著價格會降低。他們出售的數據意味著你可以使用Facebook、谷歌等真正有用的服務來改善營銷工作，而無須為此支付任何費用。然而，政府侵犯民眾隱私之後能給我們帶來什麼利益，則不是這麼直觀的。

政府的監控使我們更安全了嗎？安全機構能確保我們安然無恙嗎？的確，自2001年以來，美國就沒有發生針對平民的恐怖襲擊，至少沒有發生恐怖組織實施的恐怖襲擊。當然，對於紐約人而言，這是很有意義的。但如果僅僅根據美國沒有發生過恐怖襲擊就認為政府的安全工作做得很好，似乎沒有多少說服力，因為民眾完全有理由猜想恐怖分子再也沒有策劃過這類襲擊。政府必須讓民眾知道它挫敗了哪些襲擊計劃，不然民眾很難相信政府的話。如同得克薩斯州的沙塵暴一樣，2001年9月11日那場恐怖襲擊留給人們的記憶逐漸淡化，但在之後這麼多年間，每當討論恐怖襲擊問題時，人們總會談到由不同顏色組成的「威脅等級」示意圖，這個圖給我的感覺就像是為哈里伯頓公司精心設計的廣告一樣。政府機構往往不注重信息公開，認為民眾不需要知道任何事情，只是在有需要的時候才會向民眾透露信息。但這樣一來政府很難取得民眾的信任，而且一旦政府突然發佈某個信息，民眾可能更加關注政府為什麼會發布，而不關注信息內容。無論如何，我本人絲毫不知道美國國家安全局通過收集民眾信息減少多少犯罪，我們只是被告知他們的安全舉措有效減少了犯罪，但不知道何時、何地以及以何種方式減少了犯罪。

依靠監控攝像頭去阻止犯罪行為，無疑是堂吉訶德式的空想，但事實證明，這些監控有助於破案。^[9]如同2005年倫敦地鐵爆炸案一樣，波士頓馬拉松爆炸案的偵破工作也離不開監控數據。^[10]尤其是對於時間跨度較大的犯罪行為而言，需要獲取以往的全部數據，因為在受害者倒下之前的很長一段時間裡，罪犯就開始祕密籌劃了。在這些調查中，情報的力量頻頻見諸媒體，這種情況下，一個監視無處不在的國家的優勢就顯示出來了。監控數據有了明確的用途，受害者留在地上的血跡還斑駁可見時，也沒有人會關心保護隱私的問題，團結一致地支持政府實施監控。但在我們團結起來之前，我們對於政府行為的瞭解在很大程度上來自像斯諾登之類的洩密者。

美國國家安全局是政府收集通信情報的臂膀，他們收集的情報就在於和我們有關的各類數據中。由於個人原因，我對這個組織有一定的瞭解。正如我之前所說的那樣，我是學數學的，我在哈佛大學讀的就是這個專業。我的學士學位看起來和我同學的學位沒什麼區別，但實際上，在哈佛大學數學系有兩條不同的學習軌道，我選的這一條是為那些喜歡並擅長數學的孩子們準備的，另一條是為非常卓越的牛人準備的。本科第一年時，有一門課程被稱為「數學25」（Math 25），我學得並不好，但有一些學得非常好的學生，數學系會特別邀請他們加入一個被

稱為「數學55」（Math 55）的超級精英班。對於這些真正的數學大牛而言，我在大學期間學過的幾門最難的課程根本不值一提。在一些高級課程中承擔教學和評分任務的助教們就是這些人，他們往往比我年齡還小（一位助教甚至只有16歲），而且有的人甚至提前深入自學了研究生階段的數學課程。我記得自己當時覺得數學實分析（Real Analysis）這門課程非常難學，而我的一些同齡人——這個詞似乎也是準確的——甚至覺得這門課程非常簡單枯燥，就像九年級的課程一樣。每當我聽到美國國家安全局時，就會不由自主地想到大學時光，因為他們招的人就是來自前面所說的第二條軌道。

我之所以指出這一點，是因為在很多人看來，政府工作人員是無足輕重的，認為他們是官僚、辦事人員等。當然，在私營企業從事數據分析工作的人同樣存在不稱職的情況。然而，監視我們的人卻非常聰明，如同之前的費曼和愛因斯坦一樣，他們也都是眼光極為高遠的人，他們從事的工作都具有強大的影響力。

斯諾登透露，美國國家安全局會利用數學算法處理源源不斷收集到的數據，準確地講，是所有類別的數據。他們基本上會收集一切通過電流傳播的數據，包括電話、電子郵件、短信、照片等。顯然，這些做法不是被動的。根據一個遭到洩露的文件，政府表示其最高目標是「掌控互聯網」。^[11]這種厚顏無恥的做法非常驚人。《衛報》和《華盛頓郵報》發佈的一些幻燈片文件揭露了政府的「稜鏡」計劃。這些幻燈片直觀地說明了一切（見圖14—1）。



Gmail

facebook



Hotmail

YAHOO!

Google



skype

paltalk.com

YouTube

AOL

mail

(TS//SI//NF) PRISM Collection Details



Current Providers

- Microsoft (Hotmail, etc.)
- Google
- Yahoo!
- Facebook
- PalTalk
- YouTube
- Skype
- AOL
- Apple

What Will You Receive in Collection
(Surveillance and Stored Comms)?

It varies by provider. In general:

- E-mail
- Chat – video, voice
- Videos
- Photos
- Stored data
- VoIP
- File transfers
- Video Conferencing
- Notifications of target activity – logins, etc.
- Online Social Networking details
- **Special Requests**

Complete list and details on PRISM web page:

Go PRISMFAA

圖14—1 美國政府的「稜鏡」計劃幻燈片

「稜鏡」計劃或許應該叫「Yoink」計劃^[12]更為準確。一方面，如果一個帶著槍的人惦記著你的Facebook帳戶，那麼地球上的生活就會變得更加糟糕。另一方面，當有人悄無聲息地在微軟產品上添加了竊密工具時，人們或許根本不會注意到，更不會產生害怕情緒。

如果沒有法院的命令，任何人都無法查詢關於他人的數據，至少在理論上是這樣，因為這個計劃嚴重侵犯了個人隱私。其他竊聽的內容主要是通信的「元數據」，即發送者和接收者的電話號碼或IP地址，而不是獲取具體通信內容。下面是政府自己的「隱私和公民自由監督委員會」對另外一個計劃的部分描述：

美國國家安全局依據《愛國者法案》第215條將數百萬個電話號碼納入監控範圍，並記錄下每一個號碼的來電、去電、通話時長以及具體通話時間。當該機構鎖定某個號碼進行分析之後，與該號碼聯繫過的所有關聯號碼也會被鎖定，而且同樣的信息也會被收集，甚至與關聯號碼聯繫過的號碼也會被監控。^[13]

必須要指出的是，這些數據並不涉及通信內容。從這方面來講，這些數據與我在本書中收集的數據沒有太大的區別。我在本書中是根據用戶數據分析出整體模式，然後用這個整體模式去對應個人生活，而沒有刻意分析某一個人的具體情況，這與美國國家安全局的做法是相同的。根據「隱私和公民自由監督委員會」的說法，在美國國家安全局，只有當你的通話網絡符合他們對「威脅」的界定標準時，他們才會給予格外關注。不過，雖然他們只是獲取元數據，而非直接獲取通信內容，這並不意味著不會侵犯個人隱私。

如果有人對你很感興趣，想追蹤你，那麼你肯定會給對方留下一些令人驚訝的蛛絲馬跡。從本書前文中，你肯定已經非常瞭解這一點了。幾百年都是如此，我們沒有講到的事情還有很多。比如，Exif（可交換圖像文件）是專門為數碼相機的照片設定的，可以附加於各類圖片之中，記錄數碼照片的屬性信息和拍攝數據。從高端的單反相機到蘋果手機的內置相機，拍的照片都會自動包含這類信息，其中不僅包括照片拍攝時間，還包括攝影時的光圈、快門速度、拍攝時長、相機品牌型號以及照片拍攝者所處的經緯度等信息。正是有了Exif信息，iPhoto之類的程序才可以毫不費力地對你的照片進行排序和分類，並在地圖上標出你拍攝照片時所處的位置。Exif還會告訴你其他一些信息。以OkCupid用戶上傳的頭像為例。一張照片越好看，拍攝時間久遠的概率可能就越

大。也就是說，當人們發現一張照片不錯時，就會直接將這張照片上傳為自己的頭像，而且會使用很長時間，甚至可能永遠不會更換。我們之所以知道這一點，是因為Exif會告訴我們這張照片的拍攝時間。這類跟蹤數據是很常見的。你打開最喜歡的應用程序時，往往會有GPS座標指出你所處的地理位置。你加載的幾乎每一個網頁都有數十個單像素圖像（只是一個透明點）位於邊緣位置，真正的頁面加載時，這些圖像也會自動加載，記錄下你的訪問時間，但不會記錄你在網頁上做了什麼，只是記錄你訪問的時間以及訪問了哪些網頁。雖然這類信息很簡單，但這足以讓大數據分析公司準確地猜測到你的人口學信息。

很多人不想分享自己的信息，寧願一個人去購物，寧願自己偷著樂。這些人要怎麼做才能保護隱私呢？我本人非常清楚隱私的價值。坦率地說，我之所以不常使用社交媒體，一個原因就是為了保護隱私。我從來沒有把女兒的照片上傳到網上過，我在2011年較早的時候開始使用照片分享程序Instagram，當時該程序的服務還不是很完善，我只是將其用作一個手機相冊，因為我喜歡它的過濾器功能。我覺得它就像「創意相機」一樣，並不是一個真正意義上的社交軟件。我知道這讓我聽起來像一個老爺爺，當我妻子意識到古板的丈夫在做什麼時，她指出我可以將我的賬戶與別人的賬戶連接在一起，我照做了，因為只要點擊一個按鈕就可以了。

這種沉默是不尋常的。雖然網絡用戶很絕望，但很難說大部分用戶已經完全不在乎隱私問題了。每當Facebook更新服務條款、拓展他們的權限、日益嚴重地侵犯我們的隱私時，很多用戶總會表達憤怒，最後卻不了了之。用戶的忍耐底線一次次被拉低，就像被招惹的蜜蜂一樣，最後卻因為無人可蜇、無處可去，又不得不回到蜂巢。科技的發展會拓寬人類的認知邊界，卻不斷侵入人們的隱私，一再拉低人們的忍耐極限。目前的應用程序五花八門，有減肥的，有測心率的，有評價服裝的，你把自己的各種數據輸入進去，以期獲得時尚建議，甚至有些女性使用應用程序來預測和管理自己的月經週期。^[14]美國《紐約時報》科技撰稿人珍娜·沃瑟姆（Jenna Wortham）寫道：「市場上充斥著這類程序，我認識的幾乎每一位女性都會使用一個。」你讓應用程序知道月經開始的時間，它會提醒你排卵期。當然，雖然很多人認為自我報告的數據對個人隱私的侵犯並不是非常嚴重，但有一個新出現的應用程序卻宣稱，它可以根據女性的鏈接歷史推測出什麼時候來月經。^[15]這些與月經週期有關的應用程序背後至少有一位數據科學方面的專業人士，也能推測出懷孕、運動過度、變老或有過無保護措施的性行為，因為一旦出現這些

異常情況，用戶的登錄就會一反常態的頻繁。

儘管有些人——甚至可以說很多人——對隱私問題抱著漫不經心的態度，但本書不想透露任何個人的隱私。正如我前面所說的那樣，本書中的所有分析都是匿名的、整體性的分析。^[16]我在處理原始數據時也是謹小慎微的，數據中不會包含任何個人認證信息。我在討論中採用了許多與用戶有關的詞，包括他們的自我描述文本、在Twitter上發的帖子以及狀態更新等，但這些都是公開的，即便引用了個別用戶的發帖記錄，也對用戶賬號進行了加密。在所有分析中，引用數據的範圍僅僅侷限於基本變量，所以不會洩露任何個人隱私。

當然，我從來沒想過將數據與任何人聯繫起來。我的目標是大處著眼，分析整體情況，看看我們能夠從中獲得什麼新的認知。我認為大數據的價值就在於此，個人失去隱私的意義也在於此。《誰擁有未來？》一書的作者、微軟研究院計算機科學家杰倫·拉尼爾（Jaron Lanier）在科普雜誌《科學美國人》（*Scientific America*）上撰文指出：「與私人生活相關的海量信息在具有明確有效的用途之前，一直都在被存儲、分析和運用。」^[17]他使用的「海量」一詞無疑是正確的，但他認為信息在具有明確有效的用途之前不應該被拿去存儲、分析和運用，對此我不敢苟同。如果不這樣，怎麼知道某個事物是否有用呢？科學本身就是以探索與研究為基礎的。曾幾何時，鐵礦石長期被視為普通的石頭，直到有人開始拿它做實驗，其用途才逐漸顯示出來。在長達數百年的時間裡，麵包上的黴斑只會讓人生病，直到亞歷山大·弗萊明（Alexander Fleming）發現可以用其製作盤尼西林，其價值才得以顯現。

在數據科學領域，人們已經取得了許多意義深刻的研究成果。這些成果不只是描述人們的生活方式，還給生活方式帶來了重大的變革。比如，我在前面提到過「谷歌流感趨勢」，谷歌於2008年推出了谷歌流感趨勢服務，根據谷歌搜索數據，近乎實時地對全球當前的流感疫情進行估測，現在採用這項服務的國家超過了25個。這不是一個完美的工具，但它標誌著一個新的開端。大數據不僅僅被用來將疾病的影響降到最低，甚至還被用來預防疾病。根據2013年《紐約時報》的報道，一支由微軟、斯坦福大學和哥倫比亞大學研究人員組成的團隊，通過分析之前三年內600萬人在谷歌、微軟和雅虎搜索引擎的搜索記錄，發現了藥物潛在的副作用，他們甚至能夠在美國食品藥品監督管理局（FDA）警告系統出來之前就瞭解到處方藥的副作用。^[18]他們發現帕羅西汀（抗抑鬱藥）和普伐他汀（降膽固醇藥）同時使用會導致高血糖。從這個角度

來看，人們失去一點隱私的結果就是過上了更為健康的生活。

大數據科學領域似乎每天都會出現一些新進展，湧現一些新詞語。今天，我發現一個名為geni.com的網站已經開始嘗試為全人類創造一個在線家譜，用戶可以很方便地編輯自己的家譜樹。^[19]如果你把親人寫入族譜，系統會自動給他們發送郵件，通知並邀請他們加入該家譜，完善個人資料，並在此基礎上進一步邀請其親人加入，擴大家譜，以此類推，逐漸形成一棵標準的家譜樹。如果成功的話，該網站將勾勒出人類的基因遺傳網絡。之前的那一週，兩個政治學者顛覆了人們的一個普遍看法，即美國共和黨之所以能在眾議院奪回多數黨地位，要歸功於重新劃分選區。^[20]這兩位政治學者通過選民在互聯網上留下的搜索記錄，模擬了美國曆次選舉，得出結論，認為我們的世界本身就是分裂的，美國出現政治僵局的原因在於該國內在的政治地理特徵，而不是地圖劃分方式。

這僅僅是個開始。大數據技術一直髮展勢頭良好，一個原因是數據量非常大。比如，在2012年，Facebook每天收集500太字節的信息，數據分析技術正在迎頭趕上。^[21]納特·西爾弗使數據新聞學成了一門主流學科，給新聞報道方式帶來了深刻的變革。比如，我們理解一個問題，往往更加註重從量化角度去分析。《紐約時報》《華盛頓郵報》《衛報》都建立了強大的數據分析團隊，利用數據可視化工具製作數據新聞，即便在新聞工作者經費困難以及工作量非常大的情況下，仍然持續投入大量的資源，發佈與我們的生活有關的數據。

從蓬勃發展的大數據企業方面來看，我在本書中多次提到的谷歌公司走在最前列，利用大數據服務於公益事業，「谷歌流感趨勢」服務和谷歌公司數據科學家斯蒂芬斯—達維多維茨的工作就是如此。除此之外，該公司還推出了其他更為雄心勃勃的服務項目，但媒體宣傳比較少。比如，谷歌的數字憲法項目。這是一個以大數據為基礎的憲法庫。大多數國家的公民通常只關心一部憲法，也就是他們自己國家的憲法，但谷歌數字憲法項目卻收錄了多個國家自1787年以來制定的大約900部憲法。谷歌公司將這些憲法文本彙集在一起，進行分類和量化分析。目前，世界上每年會出現5部新憲法，而谷歌公司的數字憲法項目就會給新興國家提供有益的借鑑，使其看到歷史上什麼可行、什麼不可行，從而更有可能創建一個持久的政府。在這方面，大數據為人們開啟了更美好的未來，因為正如谷歌電子憲法項目的網站所指出的那樣，在一部憲法中，「即使一個逗號也可以產生巨大的影響」。

正如我們所看到的那樣，Facebook的數據團隊已經開始利用其龐大的數據庫研究人類行為模式與反應方式，併發布一些具有廣泛應用價值的研究成果。麻省理工學院人類動態實驗室主任亞歷克斯·彭特蘭（Alex Pentland）在牛頓學說的啟發下，將這種日漸興起的數據科學稱為「社會物理學」。^[22]他和他的團隊已經開始利用社交數據去理解和影響現實世界。他們選擇一個城市，與當地的政府、通信運營商和民眾合作，將整個城市數據化。在他們的努力下，意大利特蘭託市居民現在能夠用確鑿的數據來回答我們無法回答的問題，比如：「其他家庭的錢都是怎麼花的？」「他們將多少時間用於外出遊玩和社交？」「人們去哪些醫生那裡看病的次數最多？」

也許這就是我們的未來，這肯定是值得期待的。我在本書中分享了自己對現有數據進行創新研究的成果。我這麼做的目的並不是伸開雙臂大聲說「我攀上了研究頂峰」，而是為了讓人們領略大數據技術未來可能產生的強大能量。1953年，詹姆斯·沃森（James Watson）和弗朗西斯·克里克（Francis Crick）破解了DNA（脫氧核糖核酸）的祕密，發現了雙螺旋結構圖；60年後的今天，科學家們仍然在努力破譯和繪製人類基因圖譜。探索完整的基因表達圖譜是人類共同的事業，但在這項事業變得如此高尚之前的很多年裡，都沒有引起人類足夠的重視。

關於如何平衡大數據技術潛在的利與弊，我希望我能指出一個前進的方向。但說實話，可能是因為我長期接觸大數據，所以我認為要找到一個解決方案並非易事。我贊同拉尼爾的說法，即「監管是不可行的」。這並不是說沒有人嘗試監管。政府肯定會制定監管法律，所有的出發點也都是好的，但大數據技術發展速度太快，恐怕監管條文剛剛制定出來不久便過時了。作為一名數據收集者，我曾經給用戶提出過多種保護隱私的建議，但大多數人視若無睹，我見過很多這種情況。

OkCupid曾經向女性用戶提出過這樣一個問題：「你墮過胎嗎？」在這個問題的下面，有一個複選框，如果用戶在裡面勾選一下，其答案就會自動隱藏起來，其他用戶便無法看到。但在那些給出肯定答案的用戶裡面，勾選的用戶還不到一半。

大多數人不會使用你給他們提供的隱私保護工具，但也許這裡用「大多數人」是不合適的。一方面，為用戶提供刪除或收回數據的方法是正確之舉，無論使用者多麼少，網站管理者都應該提供。另一方面，可能人們的隱私觀念已經改變了，顯得我們這些討論隱私問題的人有點落伍。按照互聯網時代的標準，我和拉尼爾可能算是落伍的老人了。就

像在軍隊裡將軍們總是樂於打類似於上一次的戰爭一樣，我們這些人也總是固守著原有的隱私觀念。在隱私問題上，什麼是正確的，什麼是可以用於允許的，或許我的想法是錯誤的。固有文化與新一代人對隱私問題的定義是不同的。隨著時間的推移，或許我們應該突破陳規，一些老套的東西可能已經不適用了，我們需要進行革新或改變。

雖然美國國家安全局的監控行為已經嚴重侵犯了民眾隱私，但民眾似乎不那麼難過。華盛頓見證過多次百萬人大遊行，媒體報道中也會使用「上百萬男性」「上百萬母親」之類的字眼。黑客組織「匿名者」(Anonymous)曾呼籲美國民眾發起「百萬面具遊行」(Million Mask March)去反對政府的「棱鏡」計劃以及其他行為，但民眾熱情卻沒有如此高漲。在《華盛頓郵報》刊登的報道中，前幾個字卻是「數百名抗議者.....」，這直白地表明瞭民眾的冷淡反應。^[23]

拉尼爾在《科學美國人》上發表的文章建議將個人信息貨幣化，數據收集者應向網絡用戶支付小額費用。一旦人們有能力掌控自己的數據，選擇適合自己的隱私保護程度，即不同的價格，對於企業和政府而言，數據會變得過於昂貴，以至難以任意地囤積和挖掘。這一有償信息的思路將有助於個人權衡利弊，並且獲取真正的隱私權。但這種建議也存在一定的問題。如同稅收一樣，企業為了收集數據而支付的費用要麼會直接造福那些賣出自己數據的人，要麼會引發競底，導致網站不得不考慮如何用最少的成本購買最多的數據。這與航空業的競爭態勢非常相似。無論出現哪種情況，出售數據的用戶都不會得到多少價值。退一步講，這種數據貨幣化的建議也不具有現實可行性，在具體實施過程中，仍然存在著大量的細節問題，還有很長的路要走。

麻省理工學院人類動態實驗室主任亞歷克斯·彭特蘭提出的方法更具有可行性。他將自己的方法稱為「新數據協議」。頗具諷刺意味的是，他的方案背後的一些原則來源於英國古代的普通法。他認為，與任何其他的事情一樣，你對自己的數據應該享有一些最基本的權利，包括擁有權、使用權和處置權。這就意味著一旦你發現自己在某個網站或其他數據庫上的個人數據遭到了盜用，那麼你就有權從上面刪除這些數據。從理論上來講，你還應該被允許將數據據為己有，一旦形成一個合適的數據市場，你就可以轉賣自己的數據。這種簡單的機制主要是依靠網站設置完成的，網站只需要添加一個像「複製」「粘貼」按鈕那樣的「刪除」按鈕，給用戶刪除個人數據的自由。與前面提到的強制補償方案相比，該方案更具有可行性。

事實上，從企業的角度來講，我認為人們失去隱私之後已經得到了補償。他們可以使用Facebook、谷歌之類的網絡服務，聯繫上自己的老朋友，找到自己想要的東西，而這一切都是免費的。正如我前面所講的那樣，你在這些網站上輸入的個人數據越少，從中得到的服務也越少，而輸入的個人數據越多，得到的服務也就越多。人們不得不在洩露隱私與獲取服務之間做出權衡。然而，不久之後，可能用戶只需要在一個問題上做出權衡就足夠了：我要不要完全放棄這些網絡服務？因為大數據分析技術變得非常強大，可能你根本掩蓋不了自己的數據。只需要一丁點兒的個人數據，數學算法已經能夠推測出很多與你有關的事情。大數據技術才發展了短短几年就變得如此強大。不久，網站在其菜單選項中給用戶提供的一些管理隱私的設置可能根本不會起到任何保護作用，因為在這個網站以外的世界裡，不會如此謹慎地隱藏個人數據。公司和政府將能夠根據你在其他網站和其他場合留下的蛛絲馬跡發現你的存在。到那個時候，隱私或將蕩然無存，關於隱私問題的爭論也會變得不合時宜。

之前，無論我在任何情況下談到數據，我都認為數據就像洪流一般，但有一個事實或許我強調得還不夠，即數據的洪流流淌不息。只有當數據洪流停止流淌時，人們才能真正知道這條河的深度。我對此也很期待。與此同時，那些存儲、分析和運用數據的人有責任繼續採取行動，證明自己工作的價值，告訴民眾自己到底在做什麼。否則，雖然我在這本書裡為大數據技術工作者做了大量的辯護，恐怕到最後還是拉尼爾說得對，即「我們不應該這樣做」。

人們將技術視為一種新的神話。無可否認，有些技術的確具有魔力，而有的技術還存在各種缺陷。儘管如此，依然阻擋不住人們神化技術的腳步，對技術人員形成了更為高大而完美的印象，塑造出了多位「技術神靈」和技術巨人。雖然希臘的羅得島再也不像以前那麼酷了，而起源於那裡的阿波羅神像卻遍佈世界各地。同樣，雖然一些技術並不算酷，而技術神話卻不斷湧現。這就是技術行業給人的印象，悲哀的是，很多技術人員也是這麼想的。在技術領域，儘管存在一些巨人，但不存在神，我們都要牢記這一點。一切都是存在缺陷的，都是人發明的，都有一定的壽命期限，我們都生活在同一片灰暗的天空下。我們製造了數據洪流，但這股洪流是將我們淹沒，還是改善我們的處境呢？我希望我本人以及其他像我這樣的人利用數據洪流做一些真正有益的事情。在此過程中，如果我們的技術、設備以及算法看起來似乎非常好、非常先進，我們一定要回想一下丁尼生在《尤利西斯》中寫下的那句古

老的箴言，「奮鬥、探索、尋求，而不屈服」，並堅定地用一種略微不同的方法去尋找真理。我們要奮鬥、探索、尋求，最後也要懂得屈服。

[1] See Craig Smith, 「By the Numbers: 98 Amazing Facebook Stats,」 Digital Marketing Ramblings, March 13, 2014, expandedramblings.com/index.php/by-the-numbers-17amazing-facebook-stats/#.U1AArPldXko.

[2] This passage and the table are based on 「Private Traits and Attributes Are Predictable from Digital Records of Human Behavior,」 by Michal Kosinski, David Stillwell, and Thore Graepel, Proceedings of the National Academy of Sciences 110, no.15 (2013): 5802—5805.

[3] See Stephen Fairclough, 「Physiological Data Must Remain Confidential,」 Nature 505, no.7483 (2014): 263.

[4] 《自然》期刊曾經就這臺遊戲機進行過描述，指出該遊戲機配備了一個攝像頭，能夠監控到同一個房間裡的其他人的心率。感應器主要是為了運動遊戲而設計的，目的是讓玩家能夠在玩遊戲時實時監控心率變化，但這個系統也能用來監控和傳輸玩家對於電視廣告、恐怖片和政治宣傳等做出的生理反應。

[5] See David Barrett, 「One Surveillance Camera for Every 11 People in Britain, Says CCTV Survey,」 Telegraph, July 10, 2013, telegraph.co.uk/technology/10172298/.

[6] See Brian Palmer, 「Big Apple Is Watching You,」 Slate, May 3, 2010, slate.com/articles/news_and_politics/explainer/2010/05/big_apple_is_watching_you.html.

[7] cookie是瀏覽網頁後在硬盤中產生的臨時文件，它可以讓你在下次登錄該網頁時自動加載，提高網絡瀏覽速度，但也存在洩露用戶隱私的問題。——譯者注

[8] 安克誠公司宣稱：「我們成功地管理客戶群體，為客戶提供個性化的服務體驗，建立有利可圖的客戶關係。」一個有趣的悖論是，每當你看到「個性化」這個詞的時候，你大可相信結果肯定不是個性化的。

[9] See Jon Healey, 「Surveillance Cameras and the Boston Marathon Bombing,」 Los Angeles Times, April 17, 2013, articles.latimes.com/2013/apr/17/news/la-olboston-bombing-surveillance-suspects-20130417. See also 「The Need for Closed Circuit Television in Mass Transit,」 by Michael Greenberger, University of Maryland Legal Studies Research Paper No.2006—15, Law Enforcement Executive Forum (2006):151, digitalcommons.law.umaryland.edu/cgi/viewcontent.cgi?article=1065&context=fac_pubs

[10] 波士頓爆炸案發生之後，Reddit和4chan的用戶曾經努力追查肇事者，但最後卻鎖定了一個無辜的人。經過民眾的口水戰之後，還是依靠警方資源和各種硬件設施破了案。

[11] 「掌控互聯網」這個詞組特別指美國國家安全局同其他國家的情報機構開展合作，成立「五眼」情報聯盟。關於這一點，可以參考維基百科「Mastering the internet」條目。下面這個圖在《衛報》上發表之後，就迅速流傳開了，該圖的獲取鏈接為：

theguardian.com/world/interactive/2013/nov/01/prism-slides-nsa—document。

[12] Yoink指一個人將你的東西搶走時發出的得意聲音。——譯者注

[13] See David Medine et al., 「Report on the Telephone Records Program Conducted under Section 215 of the USA PATRIOT Act and on the Operations of the Foreign Intelligence Surveillance Court,」 Privacy and Civil Liberties Oversight Board (2014), <http://www.fas.org/irp/offdocs/pclob-215.pdf>.

[14] My discussion of menstruation apps is based on Jenna Wortham, 「Our Bodies, Our Apps: For the Love of Period-Trackers,」 New York Times, January 23, 2014.

[15] This fact is from Jaron Lanier, 「How Should We Think About Privacy?」 Scientific America, November 2013, 65—71.

[16] 值得再次強調的是，本書沒有任何資料透露個人隱私，至於給出的用戶照片和言論，則徵求了相關用戶的許可，具體請見相應的備註。

[17] 我對拉尼爾的討論集中在他的一篇名為「How Should We Think About Privacy?」的文章上。

[18] See John Markoff, 「Unreported Side Effects of Drugs Are Found Using Internet Search Data, Study Finds,」 New York Times, March 7, 2013, nytimes.com/2013/03/07/science/unreported-side-effects-of-drugs-found-using-internet-data-study-finds.html.

[19] geni.com宣稱已經建立了7500多萬個條目，其母公司MyHeritage宣稱建立了15億個條目。

[20] See Jowei Chen and Jonathan Rodden, 「Don't Blame the Maps,」 New York Times, January 26, 2014, nytimes.com/2014/01/26/opinion/sunday/its-the-geography-stupid.html.

[21] See Eliza Kern, 「Facebook Is Collecting Your Data—500 Terabytes a Day,」 Gigaom, August 22, 2012, gigaom.com/2012/08/22/facebook-is-collecting-your-data-500-terabytes-a-day/.

[22] My discussion of Pentland draws on his article 「Reality Mining of Mobile Communications: Toward a New Deal on Data,」 in Global Information Technology Report 2008—2009, ed. Soumitra Dutta and Irene Mia (Geneva: World Economic Forum, 2009), 75—80, and an interview with him, 「An Interview with Alex ‘Sandy’ Pentland About ‘Social Physics’」 by IDCubed: idcubed.org/?post_type=home_page_feature&p=880.

[23] See 「Million Mask March descends on Washington」 on the Washington Post's PostTV blog: <http://wapo.st/1b5Kt5J>.

後記

在涉及本書的圖表時，我參考了統計學家、藝術家愛德華·塔夫特（Edward R. Tufte）的作品。^[1]有時候不僅僅是參考，我還竭力模仿他的作品風格。他的著作言簡意賅，圖表製作精美、脈絡清晰。他在自己的作品裡提出了信息設計的一些原則。在提出這些原則之前，他借鑑了有史以來一些堪稱典範的圖表，每一個圖表都以簡略的方式呈現了豐富的數據，講述了一個故事。比如，查爾斯·米納德（Charles Minard）繪製的戰爭流程圖生動地描述了1812年拿破崙入侵俄國的情景。一位不知名的廢奴主義者繪製了一幅描述運奴船的圖，淋漓盡致地刻畫了船上擁擠不堪的情景，至今仍然被用來揭露「中間航道」^[2]的恐怖。1854年，英國倫敦霍亂暴發之後，約翰·斯諾（John Snow）博士繪製了一幅感染者居住位置分佈圖，並在人類歷史上首次指明瞭霍亂病源。塔夫特借鑑了這些畫作，並將其感悟應用於現代語境之中，指出信息設計者應該實現數據墨水比（Data-ink Ratio）的最大化，讓每一幅圖講述一個故事，用不同顏色凸顯重要信息，用白色作為一種維度的表現方式，而不只是用來表示無用的空間。我在本書圖表的繪製過程中竭力踐行他的理念。

塔夫特的著作中包含許多圖表，其中他用了兩頁的篇幅論及越南戰爭紀念碑。他並非分析紀念碑的石雕工藝或歷史價值，而是分析其在信息設計方面的價值。我希望能完整地呈現他的論述，但下面只是給出了其核心觀點：

黑色的花崗石上刻有58000名陣亡將士的姓名。從一定的距離看去，每一個姓名的字母都模糊在灰暗的背景中，但所有的名字給人帶來了最大的視覺震撼，自然讓人意識到了58000這個數字意味著什麼。

每一位數據科學家都希望發現灰暗的背景意味著什麼，我在本書中也試圖達到同樣的效果。在繪圖過程中，我著眼於最廣泛的數據，交代最多的故事背景，所有這些都旨在讓我呈現出事件真實的一面。

2008年，越南戰爭紀念碑實現了數字化，每一平方英寸都被拍了照片，並根據軍方的歷史記錄核對紀念碑上的信息，在線版的紀念碑使得訪問者能給每一個將士的名字附上照片和文本。^[3]打開網絡數據庫之

後，頁面上會給用戶提供一個空白的搜索框，讓用戶輸入關鍵字搜索陣亡將士。我稍微停頓了一下之後，便輸入了我父親的姓名戴維·巴頓·魯德爾（David Patton Rudder），因為每當想到越南，我就幾乎條件反射式地想起我父親。但很快我就想起來我父親的姓名並不在上面，這是萬幸的事。由於史密斯（Smith）這個名字用得太多，而多伊（Doe）這個名字又給人一種做作的感覺，我想了想，便輸入了約翰（John）和威爾遜（Wilson）作為關鍵詞。當然，這個名字純粹是猜測出來的。網頁只用了0.5秒，就為我搜索出來一些姓名中含有約翰和威爾遜的陣亡將士，其中我看到有這麼一位將士名叫羅恩·約翰·威爾遜（Lorne John Wilson），他的簡要信息如下：

羅恩·約翰·威爾遜

軍旅開始時間：1969—03—17

軍旅結束時間：1969—03—28

陣亡時間：1969—03—28

年齡：20歲

有人在他的條目裡附上了兩張照片，一張是身著藍色海軍制服的照片，另一張是抓拍的照片，可能是這位只做了11天海軍陸戰隊一等兵的威爾遜在國內生活的時候拍的。^[4]在這張照片上，一天下午，4個年輕人站在一輛吉普車周圍，一位站在車後方，他們正在聊天。這張照片有很多斜紋，色彩也不飽和，可能來自Instagram。這張照片的上傳者可能是他的朋友，在上傳之前肯定保存數十年了。

雖然網頁版的紀念碑無法取代用花崗石製作的紀念碑，也無法取代友誼、愛和家庭，但作為展現我們共同經歷的一個渠道，它有助於我們認識自己、認識自己的生活。現在，大數據時代已經來臨，為我們的生活留下了更為詳盡的記錄。蓬勃發展的大數據技術為我們提供了豐富多彩的服務，令我們不忍拒絕，但目前它仍然是一個政府監管盲區。如同其他類型的變革一樣，這種變革也存在令人恐懼的一面。在這二者之間，我們能找到一個合理的折中方案。我們可以藉助大數據去改進自己的認知而不操縱數據，去探索未知領域而不窺探他人的隱私，去保護大數據技術而不扼殺它，去發現而不曝光，等等。我們可以分享關於自己生活的數據，這是我們給予世界的一個無價之寶，可以為他人提供借鑑，幫助他們改善自己的人生。從古至今，人類一直在試圖實現一個共同的願望：讓自己的名字被人銘記，不僅僅留存在石頭上，還要留存在

他人記憶裡。我們分享的大數據，有利於讓每一個人實現這個最古老的願望。

[1] The discussion of the Vietnam Memorial, and the quote I use, are from *Beautiful Evidence* (Cheshire, CT: Graphics Press, 2006), but Tufte's *Envisioning Information* (Cheshire, CT: Graphics Press, 1990) and *The Visual Display of Quantitative Information* (Cheshire, CT: Graphics Press, 2001) were also indispensable.

[2] 中間航道，Middle Passage，指的是奴隸貿易船從歐洲、非洲到美洲再回到歐洲的「黑三角」航行中從非洲西海岸橫渡大西洋的那段旅程。——譯者注

[3] See fold3.com/thewall and Mallory Simon, 「Vets Pay Tribute to Fallen Comrades at Virtual Vietnam Wall,」 CNN.com, April 1, 2008, cnn.com/2008/TECH/04/01/vietnam.wall/.

[4] 海軍陸戰隊一等兵威爾遜在fold3數據庫的檔案獲取鏈接為：fold3.com/page/631972608_lorne_john_wilson/stories/。他本人是否在那張集體照裡面尚不清楚。雖然這張照片肯定拍攝於越南戰爭時期，但畫面模糊不清。

關於本書數據的說明

數據具有說服力，即便沒有上下文，數據也能讓人瞭解事實；數據清晰明確，會抑制爭論。比如，美國菸草公司在1930年投放的「好彩牌」香菸廣告中赫然寫道：「20679名醫生說好彩牌香菸的刺激性要弱一些。」但這句廣告非但無法讓我們瞭解吸菸的危害，反而會誤導我們，給我們營造出一種「吸菸無害健康」的假象，不是嗎？當數字被包裝成統計數據時，假象就會更嚴重。我不會故技重施，但每個數字背後都有一個人做出決策，包括分析什麼、排除什麼以及採用什麼樣的分析框架。發表一個聲明，即便只是做一個簡單的圖，也是一個選擇的過程。在這些選擇中，人類固有的不完美難免會表現出來。據我所知，我沒有做出任何扭曲分析結果的決定。我沒有這樣的動機，因為人類互動與交流的數據本身就已經非常有趣了，我沒有必要為了得出有趣的結論而扭曲數據。但我也做出了一些選擇，而且這些選擇也對本書產生了一定的影響，我想為你講述以下幾點。

我做的第一個選擇可能是最艱難的：在論述性和性魅力時，我決定把論述重點放到男女兩性的關係上。當然，篇幅也是我考慮的一個因素，如果將同性戀也拿來討論，就意味著必須考慮到男同性戀與女同性戀這兩種情況，使得圖表數量是當前數量的3倍。此外，一個更重要的發現是同性戀者並不例外，他們與異性戀者都表現出了同樣的傾向。比如，就像異性戀者一樣，男同性戀者也喜歡比較年輕的伴侶。對於與性間接相關的問題，比如，一個種族對另一個種族的評價，同性戀者與異性戀者也存在類似的模式。如果將論述重點放在男女兩性關係上，就能儘量避免重複，用最少的篇幅講述最多的內容，並引起最廣泛的共鳴。

我的第二個決定就是省略一些晦澀的統計學和數學概念。在做這個決定時，我的遺憾少了很多。我在本書中沒有提到置信區間、樣本量、P值以及其他類似的概念，因為這本書旨在普及數據以及數據科學。我不想費力地解釋晦澀的數學和統計學知識，但這並不能說明我的分析過程不嚴謹，就像你看不見一座房子的椽子和大梁，並不能說明房子不堅固。本書中的許多發現都源自學術性的、經過同行評議的數據源。我自己在研究過程中也應用了同樣嚴謹的標準，從一定程度上來講，還採用

了同行評議的方式，即我首先對OkCupid網站的數據進行分析，然後讓公司的一名員工去獨立驗證。此外，我將數據分析過程與數據選擇、組織過程區分開，以確保前者不會構成後者的動機。一個人負責遴選數據，另一人負責分析這些數據的意義。

有時候，我會先提出一個趨勢，然後根據自己對各種因素的理解給出一種解釋，而這種解釋通常也是我做出的最為合理的推測。對於任何一本書而言，只要其目的不是單純地羅列數據，這種做法都是有必要的。我必須從各種各樣的可能性中選擇一個最合理的解釋。我在第一章提到了「伍德森法則」，即各個年齡段的男異性戀者總是對20歲的女性最感興趣。在這個法則的背後，除了年齡因素之外，還有其他相關因素嗎？可能有，但我認為這種可能性不大。要知道，有相關性並不代表有因果關係。這句話說得非常好，我們都要牢記於心，可以促使我們將敘述重點放在真正相關的因素上，從而有效地防止過度敘述問題。但這句言簡意賅的話並不意味著因果關係的問題是無趣的，如果一種現象背後具有非常合理的原因，我會盡量列出來。

本書中的一些內容與OkCupid旗下的博客OkTrends上的帖子存在重疊現象，對於這些內容，我決定根據最新的數據重新進行分析和檢驗，而不是直接引用之前的研究成果。坦率地講，我之所以這麼做，是因為我想重新檢驗一下自己之前的分析結果是否正確。對於該博客上2009—2011年的帖子，我逐一彙集起來進行重新分析。以一個經常引用的數據為例，在這三年時間裡，多人（我至少能想起5個人）都曾經幫我統計男性與女性之間的信息回覆率。回過頭來看，我自己都不清楚他們是怎麼統計出這些數據的。重新分析一遍之後，我自己就對自己的分析結果有信心了。此外，在這個過程中，我也可以採用統一的標準，比如，我將分析侷限於20~50歲的人群。我之所以選擇這個人群，是因為我知道這個人群的代表性數據比較豐富。

由於我根據最新數據開展了新的研究，所以，本書中的數據與那個博客上的數據可能存在一些差異，曲線的彎曲方式可能略有不同，圖表顏色的深淺也可能略有不同，但本書與博客的研究結論是一致的。具有諷刺意味的是，由於這樣嚴謹的研究，如果說某些數據非常精準，反而顯得不合適，而如果說某些數據具有普遍代表性，反而較為合適。所以，我在本書中經常使用「大約」「大概」之類的字眼。當你在某篇文章中看到「89.6%的人做某件事情」時，如果運用更多的數據進行對比分析，那麼你得出的真實情況可能是「很多人」或「幾乎所有人」或「大約90%」的人做了這件事。作者之所以用了這麼精準的一個數字，

可能是因為作者認為小數點給人的感覺更酷、更權威，而下一次其他科學家給出的數字可能是85.2%，也有可能是93.4%或其他數字。你可以看著波濤洶湧的大海，問問自己究竟是哪一朵白色浪花代表著真實的海平面。你給出的答案可能是沒有意義的，甚至會誤導別人。

如果你要追溯本書的一些數據，那麼你會發現本書數據與你在OkCupid或其他網站上查到的最新數據存在差異。這是很正常的，因為這些數據與我們的生活有關，不是一成不變的，而是一直處在變化中。比如，我在寫這本書時，我的Klout影響力評分是34，而當你讀到這本書時，這個評分肯定會提高，因為我給皇冠出版社的一個承諾就是在Twitter上宣傳一下這本書，而且用戶參與度是非常高的，這簡直令人驚訝！

有時，一些數字的變化並沒有明顯的原因。比如，我和本書的編輯在谷歌搜索欄中輸入「為什麼女人」（Why do women），但谷歌給我們自動匹配的內容卻略有不同，給我匹配的內容是「穿丁字褲」，可能是因為男性經常問這個問題，而給她匹配的卻是「穿胸罩」。幾周之後，為了進一步檢驗，我在谷歌搜索欄中重新輸入了上述關鍵字，自動匹配結果卻有所不同，這一次是「穿高跟鞋」。因此，這一次是最新檢索的，所以本書中就採用了這個結果。

谷歌以及谷歌趨勢的黑色搜索框雖然非常有趣，但體現了當前數據科學領域最糟糕的特徵之一，即不透明。對於科學方法而言，驗證數據的真偽是非常重要的，但要做到這一點卻很困難，因為很多數據都是專有的。（在這方面，如同其他網站一樣，OkCupid網站也存在這種問題。）雖然大多數社交媒體大肆宣揚其掌握的數據具有多大的規模和潛力，但大部分數據都是公眾無法接觸到的。目前，在大數據研究領域，人們提起數據來源時總是謹小慎微，數據給人的感覺就像雪人一樣無跡可尋。比如，有人說「我有很多有趣的數據，但我不能說從哪兒弄到的」，有人說「我聽說天普大學的某個人在亞馬遜上獲得的評論有一大堆」，還有人說「我想L盜用了Facebook的數據」。最後這句話是我從三個互不相關的學界人士那裡聽到的，他們都說出了那個人的姓名，只是我在這裡隱去了，用L替代。事實上，L的確存在上述情況，我見過他，跟他確認過這一點，但他不會給任何人展示他的數據。其實他根本不應該持有這些數據。數據就是金錢，儲存各類數據的公司也是這麼認為的。雖然有些數據是公開的，但為了獲得這些數據，必須突破厚重的法律圍牆，而且圍牆的厚度不亞於任何金庫的牆壁厚度。如果你看了一下你朋友麗莎（Lisa）的Facebook頁面，發現她的名字是「Lisa」，並

將這個事實發布了出去，那麼無論你發佈在什麼地方，從技術上來講，你都算盜用了Facebook的數據。如果你在一個網站上註冊時提供了一個虛假的郵政編碼或虛假的生日，那麼你就違反了《計算機欺詐和濫用法案》。^[1]如果一個不滿13歲的孩子註冊或訪問《紐約時報》的網站，那就違反了該網站的服務條款，這也是違法的，不僅僅在理論上違法，實際上司法部也認定這種行為違法。我在這裡列舉的例子肯定屬於極端的情況，但法律的界定範圍都很寬，以便確保每一位使用互聯網的美國人在瀏覽網頁時都不侵犯他人的隱私權。無論你是否因自己的違法行為而遭到了懲罰，但從法律上講，你已經違法了。如果某個公司的法律顧問或某個檢察官想取悅於一個重要的企業捐贈者，那麼他們隨時可能決定起訴你，你的生活也可能由此被毀掉。如果條件合適，他們會這麼做的。因此，社會科學家在數據來源問題上向來謹小慎微。事實上，數據不僅僅像無跡可尋的雪人一樣。很多互聯網公司對其存儲的數據具有強烈的佔有慾，甚至達到了偏執的程度，而且好奇地窺探其他人或公司是否也擁有同樣的數據。

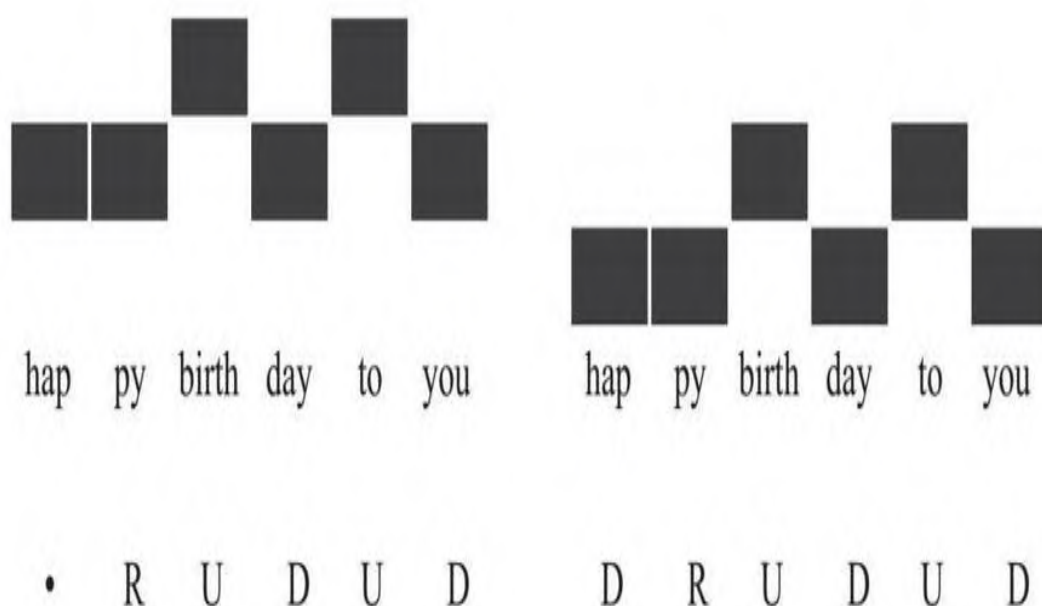
這些公司越來越喜歡把研究人員請進公司，而不是將數據公開發布出去。這種方法產生了許多成果，其中包括Facebook的數據分析團隊和谷歌公司的數據科學家斯蒂芬斯—達維多維茨有機會利用本公司的數據開展了很多新穎的研究，我在本書中也引用了他們的成果。我希望更多公司遵循這種模式，因為這樣一來，我們網站的所有者最後就能探索出一條折中的道路，既能公開發布我們的數據，促進公益事業，又不至於侵犯用戶的隱私。

∞

蘋果公司發佈的「音樂雷達」（Shazam）現在看起來似乎有點老古董的感覺，但我認為這是該公司創造的最偉大的奇蹟之一。這是一款用於識別音樂的小程序，如果你周圍在播放一首歌，你想知道這首歌是什麼，那麼你就可以打開這款軟件，將手機舉起來。這款軟件通過手機的麥克風採集聲音，只需要一兩秒的時間就能告訴你這首歌的名字和相關信息。我第一次見別人使用這個軟件時，簡直被震驚了：一方面是因為它只需要極少的信息就能識別出來，即便隔著厚厚的牆壁或處於嘈雜的環境中，它也能準確識別；另一方面是因為它的速度非常快。我一直覺得這是我見過的最神奇的事物之一。後來，我發現了它的工作原理：幾乎任何一首歌曲都可以通過旋律的高低加以識別，其他一切元素，包括基調、歌詞、韻律等，都可以忽略。要知道一首歌的名字，只需要知道

它旋律的高低就足夠了。這種旋律輪廓被稱為歌曲的「帕森斯代碼」。這簡直令人難以置信。這是音樂家帕森斯在20世紀70年代發現的。《祝你生日快樂》這首歌前兩句歌詞的帕森斯代碼就是

「·RUDUDDRUDUD」，其中U意為「旋律高」（melody up），D意為「旋律低」（melody down），R意為「重複音」（repeated note），前面那個點表示歌詞的開始，當然也就談不上旋律的高低了。這兩句歌詞的帕森斯代碼用圖案表示如下，你也可以哼唱一下這兩句，來驗證一下：



非常神奇的是，不僅這兩句歌詞的帕森斯代碼是獨一無二的，幾乎所有音樂的代碼都是獨一無二的。由於這些代碼採用的字母簡單明瞭，所以「音樂雷達」可以快速識別。以保羅·麥卡特尼的《昨日》

（*Yesterday*）為例。這款軟件不需要知道這是一首吉他演奏的音樂，也不需要知道它的演唱者，只需要收集到少量的歌聲，根據這首歌第一個單詞*Yesterday*的帕森斯代碼「·DRUUUUUDDR」就能識別出這首歌。這是比較容易理解的。

如同這款應用程序尋找歌曲的旋律模式一樣，數據科學的目標也是尋找模式。我和其他從事類似工作的同行為了在噪聲中找到信號，一次又一次地設計方法和結構，甚至想方設法探索出一條捷徑。我們都在尋找自己的帕森斯代碼。這項工作看似簡簡單單，實際上卻非常艱難，即

便一生只能取得一次重大的發現，也已經算是幸運了。雖然數據科學可能面臨這樣或那樣的問題，雖然失敗的概率非常大，但我通過本書表達了自己的觀點，即我喜歡這項事業。

[1] 如果想進一步瞭解《計算機欺詐和濫用法案》的嚴苛規定，請參考「電子前線基金會」刊登的兩篇文章：一篇是《直到今天，如果你17歲就瀏覽seventeen.com，那就違反了〈計算機欺詐和濫用法案〉》（Until Today, If You Were 17, It Could Have Been Illegal to Read Seventeen.com Under the CFAA），另一篇是《你是一位在線閱讀新聞的青少年嗎？根據司法部的說法，你可能違法了》（Are You a Teenager Who Reads News Online? According to the Justice Department, You May Be a Criminal）。

致謝

首先，我要感謝我的妻子萊西瑪持之以恆的支持和無私的愛。如果沒有她，這本書以及我的人生恐怕就像沒有裝訂起來的書頁一樣，被風吹得凌亂不堪。

感謝馬克斯·克羅恩、薩姆·亞甘和克里斯·科因協助我建立和經營OkCupid網站，在過去15年間，與你們一路走來，我深感榮幸。

感謝我的經紀人克里斯·帕里斯—蘭博，我曾經在酒吧裡和他聊過寫這本書，是他給我提供了很多切實可行的建議，最後才有了這本書。感謝皇冠出版社的編輯阿曼達·庫克。她很有耐心，也很有編輯技巧，讓一些枯燥的理念變得頗具可讀性，可以說，這本書的成功與她密不可分。我還要感謝編輯助理艾瑪·貝利和設計團隊——尤其是克里斯·布蘭德，她們為本書付梓付出了很多努力。我也要感謝安斯利·羅斯納、薩拉·布里沃喬和薩拉·裴克代米爾以及傑伊·鬆斯幫助這本書走向世界。莫莉·斯特恩、雅各布·劉易斯和戴維·德拉克的支持和遠見讓上述這些努力變成了現實。我還要感謝企鵝出版社的阿莉森·洛倫森很早就對本書的出版事宜提供了指導。

感謝我那堪稱「多面手」的數據研究員和程序設計師詹姆斯·多德爾，他建立的數據庫對於本書寫作具有不可替代的作用，本書的很多地圖與網絡圖表都是他製作的。感謝湯姆·奎塞爾和邁克·馬克西姆多次幫我從OkCupid網站上遴選數據，並在數據統計方面給我提出了非常好的建議。

感謝我的父母和妹妹給我的鼓勵，你們是我生命的根基。感謝佩特爾一家人支持我和雷什馬日復一日地撰寫這本書。

感謝一些供職於各大網站的朋友幫助我整合與獲取數據，其中包括Shiftgig的艾迪·盧、StumbleUpon的蒂姆·亞伯拉罕（現在供職於Twitter）、Tinder的萊恩·奧格爾和肖恩·萊德、Match的吉姆·塔爾博特，Datehookup的湯姆·雅克、Reddit的埃裡克·馬丁。感謝邁克爾·塔珀和本·穆雷幫我審讀初稿。感謝Mathey & Tree律師事務所的肖恩·馬修、弗蘭克林·韋恩里布、Rudell & Vassallo律師事務所的埃裡克·布朗以及Smith

Anderson律師事務所的約翰·泰瑞恩為我提供法律方面的協助。

感謝道格·德梅為我提供的建議，雖然是以非正式的形式提出的，但建議很好。最後，我要感謝傑德·麥克萊布和賈斯汀·萊斯，從比特幣到歌手鮑勃·迪倫再到尤利西斯，你們教我懂得了很多。你們的友情讓我的生命與這本書更加充實。

圖書在版編目（CIP）數據

無人旁觀時我們是誰：大數據下的人類真實面目/（美）克里斯蒂安·魯德爾著；蔣宗強譯.--北京：中信出版社，2020.12

書名原文：Dataclysm: Who We Are (When We Think No One's Looking)

ISBN 978-7-5217-2073-0

I .①無... II.①克...②蔣... III.①心理交往—社會心理學 IV.①C912.11

中國版本圖書館CIP數據核字（2020）第145626號

無人旁觀時我們是誰：大數據下的人類真實面目

著者：〔美〕克里斯蒂安·魯德爾

譯者：蔣宗強

出版發行：中信出版集團股份有限公司

（北京市朝陽區惠新東街甲4號富盛大廈2座 郵編 100029）

承印者：

開本：787mm×1092mm 1/16

印張：22

字數：221千字

版次：2020年12月第1版

印次：2020年12月第1次印刷

京權圖字：01-2013-7136

書號：ISBN 978-7-5217-2073-0

定價：68.00元

版權所有，侵權必究

Table of Contents

[作品介紹](#)

[書名頁](#)

[目錄](#)

[前言](#)

[第一部分 我們因何而聚](#)

[第一章 伍德森法則](#)

[第二章 出醜效應](#)

[第三章 “作家”的世界](#)

[第四章 社交圖譜](#)

[第五章 “約會大冒險”：雖敗猶榮](#)

[第二部分 我們的隔閡從何而來](#)

[第六章 混淆變量](#)

[第七章 被神化的美貌](#)

[第八章 隱祕的選擇](#)

[第九章 憤怒的時代](#)

[第三部分 影響身份認同的因素](#)

[第十章 你是誰？](#)

[第十一章 你墜入愛河了嗎？](#)

[第十二章 瞭解自己所處的位置](#)

[第十三章 個人品牌](#)

[第十四章 蛛絲馬跡](#)

[後記](#)

[關於本書數據的說明](#)

[致謝](#)

[版權信息](#)