

資訊保安不善 大數據難起飛

作者：香港中文大學工程學院副院長（外務）、香港資訊科技聯會前會長 黃錦輝

一九六二年颱風「溫黛」襲港，造成近六百人傷亡。五十年後，颱風「海鷗」同樣來勢洶洶吹襲香港，不同的是「海鷗」並沒有構成重大傷亡。市民能避過無情暴風的蹂躪，筆者認為香港天文台居功至偉。近年天文台引入先進科技監察氣象變化，從大量氣象數據中進行大規模的實時分析，提供精準及適時的預警及通告，令市民預早做好防風措施，確保自身及家居安全。這例子反映出「大數據」分析對社會的重要性。

更新率異常迅速

另一例子是美國總統奧巴馬深信資訊科技的重要性，在第一任期時成功游說國會撥款二十億美元推動全國「大數據分析計畫」，目的是鼓勵聯邦政府內部及各省市進行跨部門數據交流及分析，使政府能徹底了解市民的需求，繼而為市民提供更多、更好兼度身訂造的服務。簡單而言，「大數據」是指非比尋常、極度複雜的海量數據，其質與量已超越現有數據庫管理技術的處理能力。國際資訊科技專家，於初期用三個 V 來描述大數據的技術特徵：

大數據更新率（Velocity）異常迅速。例如現時全球各財務機構每秒處理一萬個以上的信用卡交易；美國沃爾瑪公司每小時處理超過一百萬名用戶的買賣；全世界超過五百萬人利用智能移動電話通話、上網、發短信、玩網遊等。這些應用與我們的日常生活息息相關，是典型的「大數據」源頭。據資訊科技專家估計，大數據容量（Volume）在未來十年，比現時的數據量將超越五十倍，估計其容量將會超過「堯字節」（yotta-bytes, YB, 24 位數目）之多。除了超海量之外，未來大數據的結構種類（Variety）將變得五花八門，除了數、文、圖、音、錄像等多媒體數據之外，也不乏社交網絡的關係圖（Social Network Graph）、遙感器訊息（Sensor Data）、半結構化數據（Semi-Structured Data）、知識（Knowledge）、規則（Rules）等非傳統數據。總括而言，在未來十年全球 IT 業界必須解決因「大數據」的容量（Volume）、異構性（Variety）及更新率（Velocity）三者所產生的一系列技術難題。

網絡入侵影響信心

然而，單是從技術層面不能全面顯示大數據的特徵，業界進一步從應用角度加插了另外兩個 V。工程師利用先進軟件及統計技術來分析大數據，從中挖掘出有價值（Value）的知識，應用系統因而變得更優化、更高效。可是「垃圾進，垃圾出」（Garbage In Garbage Out, GIGO），不管分析技術如何先進，若果輸入的數據缺乏直實性（Veracity），輸出的結果必然不可靠。倘若不可靠的結果最終被採用，情況恐怕只會弄巧反拙。

筆者認為上述五個 V 還未足夠，未能確保大數據的實用性（practicality）。近期網絡入侵以至資訊盜竊及資訊毀壞等網絡罪行於全球猖獗，大數據庫（尤其是集中式雲端儲存伺服器）存在極大的保安穩患（Vulnerability）危機。這現象不但影響數據質量之餘，亦嚴重影響用戶對大數據應用的信心。網絡入侵事件比比皆是，最為全球矚目的是二〇一三年五月的美國真人版「竊聽風雲」事件，美國前國安局技術員斯諾登披露美國政府在外國進行資訊盜竊活動，受害者包括德國總理默克爾。近期最熱烘烘的案例是美國「艷照門」醜聞，事件中多個國際知名女星的自拍裸照，從雲端儲存器 iCloud 中被盜取。由此可見，隨着社交網絡服務的普及化，私隱盜竊的情況會愈來愈常見。

所謂「一年被蛇咬，十年怕草繩」，以德國為例，竊聽事件過後默克爾便計畫與歐盟諸國設置歐洲內聯網，防止美國再次竊聽。筆者認為這方案太消極，治標不治本，不但損害了網絡的自由度及資訊的流通性，亦影響世界宏觀經濟發展。政府及專家們必需要對症下藥，資訊保安問題一日未解決，大數據產業難有太大的發展空間。