STUDIA PRZYPADKÓW 9

Prezentując w tym rozdziale wybrane projekty dotyczące analizy sieci złożonych, pragniemy uzmysłowić Czytelnikom, że ta dziedzina, choć intensywnie opisywana od półwiecza, wciąż pozostaje otwarta na nowe pomysły i zastosowania. Przełomowe koncepcje pojawiały się w niej wskutek pozyskania nowych danych albo dostępności nowych technologii ich przetwarzania. Nic nie wskazuje na to, żeby w przyszłości miało zabraknąć jednych albo drugich.

Prezentację rozpoczynamy od zagadnienia ustalania tożsamości węzłów. Następnie przechodzimy do omówienia wyników własnych prac dotyczących badania niezawodności polskiej sieci internet, wykrywania ataków sieciowych i modelowania procesów migracji klientów sieci komórkowej. Kolejny temat dotyczący dynamiki sieci przedstawia mechanizmy przemiany triad w serwisie społecznościowym. Naszą antologię zamyka zagadnienie komplementarne do otwierającego, której przedmiotem jest anonimizacja tożsamości węzłów.

9.1. Deanonimizacja sieci

Każda sieć stanowi zbiór danych. Rodzi to skutki prawne dla jej posiadacza związane z ich ochroną zarówno we własnym interesie, jak również w interesie podmiotów, których te dane dotyczą. Z drugiej strony, istnieje całkiem długa lista pokus udostępnienia *niektórych* z posiadanych danych podmiotom zewnętrznym. Dane tego typu stanowią obecnie cenny towar dla różnych klientów – banków, firm ubezpieczeniowych, podmiotów publicznych, agencji reklamowych itp. Posiadacz danych może również chcieć je wydać w celu ich opracowania na własne potrzeby przez instytucję zewnętrzną –

uczelnię, firmę doradczą – w przypadku, gdy sam nie dysponuje odpowiednimi umiejętnościami analitycznymi.

W celu ochrony indywidualnych danych związanych z poszczególnymi węzłami sieci przeprowadza się anonimizację sieci przed jej wydaniem na zewnątrz. Konkretna technika anonimizacji jest dostosowana do sposobu wykorzystania sieci przez podmiot zewnętrzny i wynika z konieczności ochrony danych osobowych i tajemnicy biznesowej właściciela oraz potrzeb podmiotu zewnętrznego. Do realizacji niektórych celów wystarczy udostępnienie danych statystycznych pozbawionych informacji o strukturze powiązań w sieci. Gdy jednak decydujemy się na udostępnienie tej struktury, pozbawionej tożsamości węzłów, należy mieć świadomość, że istnieją metody rekonstrukcji ich tożsamości.

W ogólności, przedmiotem deanonimizacji jest sieć $G = (V, E)$ pozbawiona informacji o tożsamości węzłów – z wyjątkiem kilku z nich. Oznacza to, że osobnik zainteresowany deanonimizacją (nazwijmy go umownie włamywaczem) zna wstępnie tożsamość kilku węzłów. Oznaczmy tę wiedzę jako zbiór par: (węzeł, tożsamość):

$$M = \{(v_i, \theta_i), (v_j, \theta_j), \dots\}, \tag{9.1}$$

gdzie $\theta_i, \theta_j, \dots \in \Theta$ to identyfikatory tożsamości ze znanego włamywaczowi zbioru Θ. Początkową tożsamość węzłów można poznać, wyszukując te, które same ją nierozważnie udostępniają, albo skłaniając je w różny sposób do jej udostępnienia. Są to metody pasywne pozyskiwania tożsamości, gdyż nie ingerują w strukturę sieci. Alternatywnie włamywacz może wprowadzić do sieci własne węzły tworzące unikatową strukturę, tak aby można ją było następnie odnaleźć w sieci zanonimizowanej (tzw. *Sybil attack*).

Poniżej przedstawiamy strategię deanonimizacji opisaną w [111]. Jej autorzy przyjęli, że oprócz sieci anonimowej jest znana inna sieć, $G^+ = (V^+, E^+)$, o jawnej tożsamości wierzchołków. Wykorzystali w tym celu sieć połączeń portalu Flickr. Za sieć zanonimizowaną przyjęli powiązania między użytkownikami usługi Twitter. Zatem $\Theta \equiv V^+$.

Struktury G i G^+ nie są, rzecz jasna, jednakowe – dlatego zaproponowana metoda deanonimizacji poszukuje par węzłów (v_i, v_i^+) o najbardziej zbliżonej strukturze powiązań. Szczegóły działania algorytmu są przedstawione w postaci pseudokodu – algorytm 9.1, a wyniki działania na przykładowej parze grafów – rys. 9.1. Początkowo, dzięki M, można zidentyfikować dwa tożsame podgrafy, H i H^+ (wiersz 1 wydruku). Kojarzenie kolejnych par węzłów odbywa się iteracyjnie, dla kolejnych węzłów-sąsiadów H. Każda iteracja (wiersze 2–19) rozpoczyna się od wytypowania węzła anonimowego v_k dobrze połączonego ze zbiorem węzłów już zdeanonimizowanych (wiersz 3), a następnie wyszukania w sieci jawnej dobrego odpowiednika (wiersze 4–7), tj. połączonego z H^+ w sposób zdecydowanie odróżniający go od pozostałych. Jeśli taki potencjalny odpowiednik, v_l^+, istnieje, to prowadzi się analogiczne poszukiwanie dla odwzorowania odwrotnego, tj. w sieci G (wiersze 8–11). Jeśli wynikiem tego poszukiwania jest wyjściowy węzeł v_k, to odwzorowanie uznajemy za udane, uzupełniamy zbiór wiedzy M i dodajemy utożsamione węzły do grafów H i H^+ (wiersze 12–13).