Algorithms for analyzing evolving networks on the Dark Web & in science

Boekhout, H.D.

Samenvatting

Grootschalige data verzameling en analyse is in toenemende mate aan de orde van de dag. Traditionele data analyse beschouwt informatie met betrekking tot elk individueel object en/of entiteit in een systeem los van elkaar. Netwerk analyse daarentegen, bestudeerd de interactie tussen de objecten/entiteiten, door deze te modelleren als *knopen* en de knopen te verbinden met links (*takken*) als interactie plaats vindt. Dit proefschrift bestudeerd twee "real-world" systemen met behulp van netwerk analyses waarin tijd, i.e., het moment van interactie, wordt meegewogen. Door het moment van interactie mee te wegen, kan een dergelijke analyse tot rijkere inzichten leiden. Er bestaat echter geen eenduidige manier om temporele informatie te modelleren in een netwerk en moet er dus voor elke analyse overwogen worden wat de beste methode is om het model zo goed mogelijk aan te laten sluiten bij zowel de realiteit als de onderzoeksvraag. Tevens bestaan er niet altijd (efficiënte) temporele varianten van algoritmen om de benodigde analyse te kunnen uitvoeren. Dit proefschrift introduceert daarom ook nieuwe (efficiënte) algoritmen ter ondersteuning van de analyses.

Het eerste systeem dat wordt bestudeerd, is de "Dark Web cryptomarket" EVOLUTION. EVOLUTION was actief van januari 2014 tot maart 2015, en was een combinatie van forum en marktplaats waar anoniem illegale goederen en diensten (bijv. drugs, ID, credit- en debitcard informatie, of wapens) konden worden verhandeld en bediscussieerd. In hoofdstuk 2 wordt de ruwe data, die over deze cryptomarket was verzameld en gepubliceerd in de *Dark Net Market archives*, bestudeerd. Hierbij worden data kwaliteit problemen zoveel mogelijk geïndentificeerd en opgelost, de mate van missende data geschat, en de forum en markt data aan elkaar gelinkt. (Temporele) data inconsistenties blijken vaak makkelijk opgelost of daadwerkelijke veranderingen te

representeren, en dubbele data blijkt vaak gebruikt te kunnen worden om ontbrekende informatie aan te vullen. Zodoende kan het raadzaam zijn om bij het verzamelen van data via "scraping" meermaals dezelfde data, i.e., webpagina's, te verzamelen.

Hoofdstuk 3 maakt vervolgens gebruik van de opgeschoonde data om te bestuderen of communicatie op het forum een voorspellende waarde kan hebben op het aantal verkochte producten en/of diensten van verkopers op de markt. Zowel traditionele data analyse methoden, zogenaamde activiteit indicatoren, als netwerk analyse methoden, zoals centraliteitsmaten, worden met elkaar vergeleken. Met behulp van temporele informatie over wanneer communicatie op het forum plaatsvond als wel wanneer verkopers bepaalde verkoopresultaten bereikten, kan niet alleen iets gezegd worden over de voorspellende waarde voor huidig verkoopsucces, maar ook toekomstig succes. Voor de netwerk analyses worden forum gebruikers bovendien alleen verbonden mits de communicatie binnen een voldoende korte tijd van elkaar plaatsvond. Zodoende wordt de temporele informatie gebruikt om de modellering van het systeem dichter bij de realiteit te brengen en beter te laten aansluiten bij de onderzoeksvraag. Tegelijkertijd zijn dankzij deze aanpak geen temporele algoritmen nodig voor de analyse. *Topic engagement*, een maat van hoe vaak gereageerd wordt op topics gestart door de gebruiker, blijkt de beste indicator van huidig en toekomstig verkoopsucces. De centraliteitsmaat *betweenness centrality* blijkt juist beter in het vinden van succesvolle verkopers die relatief weinig actief zijn op het forum. Hiermee speelt betweenness centrality goed in op een zwakte van topic engagement. Bovendien blijkt een hoge betweenness centrality een goede indicator voor gebruikers met centrale rollen, zoals administrator, die belangrijk zijn voor het functioneren van zowel forum als markt.

Het tweede systeem dat wordt bestudeerd, is het wetenschappelijke publicatie systeem. Specifiek wordt in Hoofdstuk 5 gekeken naar de evolutie van het netwerk van samenwerkingen op publicaties tussen onderzoekers in verschillende steden, en hoe dit verschilt van simulaties. Gebaseerd op de observatie dat wetenschappelijke adressen in de grote steden vaak geclusterd bij elkaar liggen, worden in dit hoofdstuk *scientific cities*, clusters van wetenschappelijke adressen, geïntroduceerd als een nieuwe ruimtelijke eenheid. Deze eenheid kan eerlijker tussen steden vergelijken wanneer we aannemen dat het ongeveer even makkelijk is (qua reistijd) om samenwerkingen aan te gaan binnen elke stad. Gezien reistijden binnen een wereldstad als New York veel langer kunnen zijn dan in een kleine stad als Leiden, voldoen traditionele stadsdefinities immers niet aan deze aanname. Samenwerkingen tussen scientific cities kunnen dan ook gezien worden als samenwerkingen op langere afstanden. Afstanden waar digitaal overleg al snel makkelijker wordt dan face-to-face.

Hoofdstuk 5 beschouwt dus in essentie het netwerk van samenwerkingen tussen wetenschappers op afstand. De evolutie van dit netwerk wordt bestudeerd door het netwerk op te splitsen met "time-slices", oftewel door de volledige periode van data

op te splitsen in meerdere korte perioden, en de resulterende netwerken met elkaar te vergelijken. Gebaseerd op de rankings van meerdere centraliteitsmaten, blijkt dat het netwerk over de jaren meer stabiel is geworden. Om de evolutie van het netwerk over de tijd te vergelijken met een gesimuleerde evolutie, wordt er een nieuw EDRR netwerkmodel geïntroduceerd. Waar synthetische netwerken vaak vanuit het niets worden opgebouwd, worden EDRR netwerken gegenereerd met als basis het "real-world" netwerk van de voorgaande tijdstap. Vervolgens worden random een gelijk aantal tak mutaties toegepast als dat in de "real-world" netwerken plaatsvond. Vergelijkingen van de "real-world" evolutie met gesimuleerde EDRR netwerken impliceren dat nieuwe samenwerkingen tussen scientific cities vaker plaatsvinden tussen steden die niet ver uit elkaar liggen in de periferie van het netwerk.

Hoofdstuk 6 bestudeert vervolgens wetenschappelijke teams die langdurig samenwerken. Deze langdurige samenwerkingen worden bepaald gebaseerd op temporele cliques, groepen van volledig verbonden knopen, in het wetenschappelijke samenwerking netwerk. In hoofdstuk 4 wordt hiervoor een algoritme geïntroduceerd om deze temporele cliques veel efficiënter te bepalen dan met bestaande algoritmen. Deze efficiëntie wordt bereikt onder andere door middel van het efficiënt snoeien van de zoekboom. Om het algoritme te evalueren wordt, in hoofdstuk 4, tevens ook een nieuwe methode geïntroduceerd om synthetische temporele netwerken te genereren die variëren in structurele en temporele dichtheid.

Teams met langdurige samenwerkingen blijken betrokken te zijn bij het overgrote deel van de wetenschappelijke publicaties en in overmaat bij succesvolle publicaties met veel citaties. Bovendien ondervinden deze teams vaker vroeger dan later in hun samenwerking hun succes, met de grootste succeskansen in de eerste twee jaar. Dit gaat in tegen de gebruikelijke opvatting dat teams succesvoller worden naarmate ze over de tijd beter op elkaar ingespeeld raken. Opvallend is dat als dit soort teams ontstaan vanuit bestaande samenwerkingen, dat het nieuwe team veel vaker vroeg succes behaald. Er is dus wel degelijk waarde in beter leren samenwerken over de tijd, er is alleen vaak een nieuwe frisse impuls nodig voor succes. Dit zien we ook later in de samenwerking terug, waar als teamleden vaker (maar niet te vaak) betrokken zijn bij het ontstaan van nieuwe samenwerkingen, de kansen van succes van het team zelf ook stijgen. Zodoende worden langdurige samenwerkingen niet afgeraden, zolang dit niet in de weg staat van het aangaan van nieuwe, frisse samenwerkingen.

Centraal in deze thesis staat het bij netwerk analyse op de juiste manier omgaan met temporele informatie. In deze thesis hebben we twee manieren gezien om dit aan te pakken: (1) de temporele data gebruiken om het systeem te modelleren als meerdere netwerken, elk met een eigen tijdspanne, waarna niet-temporele methoden toegepast kunnen worden (hoofdstukken 3 en 5); en (2) de temporele data als onderdeel van het netwerk modelleren en temporele methoden toepassen (hoofdstukken 4

en 6). Dit laat zien dat de keuze meestal niet alleen afhangt van het systeem, i.e., de onderliggende data, dat wordt gemodelleerd, maar in grote mate afhangt van de onderzoeksvraag. Daarnaast, blijken de juiste methoden, om synthetische data te genereren voor vergelijking met "real-world" data in de temporele setting, vaak nog niet te bestaan. De ontwikkeling hiervan lijkt echter wel meestal op niet-temporele modellen voort te kunnen bouwen.